

BRNO UNIVERSITY OF TECHNOLOGY

Faculty of Electrical Engineering
and Communication

BACHELOR'S THESIS

Brno, 2020

Viktória Parobková



BRNO UNIVERSITY OF TECHNOLOGY

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

FAKULTA ELEKTROTECHNIKY
A KOMUNIKAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF BIOMEDICAL ENGINEERING

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

BIOINFORMATIC ANALYSIS OF SINGLE NUCLEOTIDE POLYMORPHISMS IN THE 1000 GENOMES PROJECT DATABASE

BIOINFORMATICKÁ ANALÝZA JEDNONUKLEOTIDOVÝCH POLYMORFISMŮ V DATABÁZI 1000
GENOMES PROJECT

BACHELOR'S THESIS

BAKALÁŘSKÁ PRÁCE

AUTHOR

AUTOR PRÁCE

Viktória Parobková

SUPERVISOR

VEDOUCÍ PRÁCE

prof. Ing. Ivo Provazník, Ph.D.

BRNO 2020

Bachelor's Thesis

Bachelor's study field **Biomedical Technology and Bioinformatics**

Department of Biomedical Engineering

Student: Viktória Parobková

ID: 203681

**Year of
study:** 3

Academic year: 2019/20

TITLE OF THESIS:

Bioinformatic analysis of single nucleotide polymorphisms in the 1000 Genomes Project database

INSTRUCTION:

1) Study the whole genome mapping of DNA sequence variations (single nucleotide polymorphisms, short deletions and insertions). Include results of genome mapping of individuals and mother-father-child trios into the research. 2) Study the mapped genomic structural variants described by allele frequencies and correlation patterns between nearby variants. 3) Understand the structure of 1000 Genomes project database, design software for reading and processing data from these databases. 4) Perform bioinformatic, statistical, and population analysis of DNA variation data. 5) Present and discuss the results of the analysis appropriately. The preferred language for the project is English.

RECOMMENDED LITERATURE:

[1] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 467:1061-1073, 2010.

[2] International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* 467: 52-58, 2010.

**Date of project
specification:** 3.2.2020

Deadline for submission: 5.6.2020

Supervisor: prof. Ing. Ivo Provazník, Ph.D.

prof. Ing. Ivo Provazník, Ph.D.
Subject Council chairman

WARNING:

The author of the Bachelor's Thesis claims that by creating this thesis he/she did not infringe the rights of third persons and the personal and/or property rights of third persons were not subjected to derogatory treatment. The author is fully aware of the legal consequences of an infringement of provisions as per Section 11 and following of Act No 121/2000 Coll. on copyright and rights related to copyright and on amendments to some other laws (the Copyright Act) in the wording of subsequent directives including the possible criminal consequences as resulting from provisions of Part 2, Chapter VI, Article 4 of Criminal Code 40/2009 Coll.

ABSTRACT

The whole-Genome sequencing and its variations discovery were challenging for many years. The knowledge of all genetic variations is remarkably beneficial in disease research. The bachelor thesis is dedicated to human genetic variations and its two main research projects, the HapMap Project and the 1000 Genomes Project which notably helped the disease analyses. The first part describes both Projects and the following part explains the structure of their databases and presents software which enables to browse and download data from these projects. At last, statistical, population and bioinformatic analyses are performed on structural variant dataset assembled by the 1000 Genomes Project.

KEYWORDS

DNA variations, polymorphism, haplotype, HapMap Project, 1000 Genomes Project, genotyping, sequencing, VCF, gene

ABSTRAKT

Sekvenovanie celého ľudského genómu a nájdenie jeho variácií bolo výzvou počas mnohých rokov. Znalosť všetkých genetických variácií je pozoruhodne prospešná pri výskume chorôb. Táto práca je zameraná na genetické variácie človeka a ich dva hlavné výskumné projekty, The HapMap Project a The 1000 Genomes Project, ktoré pomohli v analýze chorôb. Prvá časť práce je venovaná teoretickému popisu projektov. V nasledujúcej časti práce sú popísané štruktúry databáz u oboch projektov a taktiež je predstavený online nástroj umožňujúci prehľadávanie a sťahovanie ich dát. Následne je prevedená štatistická, populačná a bioinformatická analýza štrukturálnych variácií produkovaných 3 fázou 1000 Genomes projektu.

KĽÚČOVÉ SLOVÁ

DNA variácie, polymorfizmus, haplotyp, HapMap Project, 1000 Genomes Project, genotypovanie, sekvenovanie, VCF, gén

PAROBKOVÁ, Viktória. *Bioinformatic analysis of single nucleotide polymorphisms in the 1000 Genomes Project database*. Brno, 2020, 62 p. Bachelor's Thesis. Brno University of Technology, Faculty of Electrical Engineering and Communication, Department of Biomedical Engineering. Advised by prof. Ing. Ivo Provazník, Ph.D.

ROZŠÍRENÝ ABSTRAKT

Celogenómové sekvenovanie sa považuje za jeden z hlavných progresov v bioinformatike, molekulárnej biológii a mnohých iných odvetviach. Prinieslo možnosť lepšieho pochopenia častých aj neobvyklých ochorení z ich genetického hľadiska. Osekvenovanie celého genómu trvalo roky a podieľalo sa na ňom viacero výskumných skupín z celého sveta. Najväčší úspech sekvenovania bola platforma Illumina, ktorá dokázala osekvenovať celý ľudský genóm za 50 hodín, s tým že samotné sekvenovanie trvalo necelých 20 hodín. Možnosť osekvenovania celého genómu pomohla v hľadaní a genotypovaní variácií v ľudskom genóme, ktoré môžu byť genetickou príčinou ochorenia.

Začiatok bakalárskej práce je venovaný základom genetiky a postupom detekovania variácií v ľudskom genóme. V ďalších kapitolách sa práca zaoberá dvomi projektami, ktoré sú spojené s objavovaním a analýzou variácií, a sú to The HapMap Project a The 1000 Genomes Project. Prvé dva body zadania bakalárskej práce sú prevažne obsiahnuté v rešerši oboch Projektov, keďže celogenómové mapovanie variácií DNA, a genomické štrukturálne varianty študované podľa alelových frekvencií a korelačných modelov boli uskutočnené Projektami.

Cielom HapMap projektu bolo katalogizovať jednonukleotidové polymorfizmy, známe ako SNP, v ľudskom genóme a umožniť vedcom rýchlejšiu a lacnejšiu analýzu ochorení. Ideou bolo použitie tag SNP, ktoré reprezentujú ostatné SNPs v haplotype, ktoré sú v kompletnej väzbovej nerovnováhe (anglicky linkage disequilibrium LD) a tak genotypovať iba 200,000 až 1,000,000 tag SNPs namiesto 10 miliónov SNPs, potrebných pre štúdiu celého genómu. Dáta sú voľne dostupné pre celú verejnosť v databáze na FTP stránkach. The 1000 Genomes Project študoval nie len SNP ale aj iné typy variant, ako krátke delécie, duplikácie, štrukturálne varianty a variability počtu kópií (anglicky copy number polymorphism CNP). Vzorky pre štúdiu pochádzali z rôznych populácií a kontinentov a cieľom bolo vytvoriť katalóg genetických variácií obsahujúci aj variácie s frekvenciou pod 1%. Dáta sú rovnako ako pri HapMap projekte verejne dostupné napríklad na *NCBI (The National Center for Biotechnology Information) FTP stránke* alebo *1000 Genomes FTP stránke*. Štruktúra databáz oboch projektov je upresnená v kapitole 6.2. Jeden z možných spôsobov ako vyhľadávať dáta publikované projektami je použitie online nástroja The 1000 Genomes Browser sprostredkovaného NCBI stránkou. V zadaných oblastiach je možné vidieť nájdené varianty spolu s ich alelovou frekvenciou pre individuálne populácie alebo dokonca aj pre individuálne vzorky. Tieto informácie sú voľne stiahnuteľné a použiteľné pre rôzne štúdie.

V programovej časti bakalárskej práce bola prevedená bioinformatická, štatistická a populačná analýza DNA variácií z 1000 Genomes Project v jazyku R. Výber použitých metód a ich postup vychádza z *publikácie* tretej fázy 1000 Genomes

Project, kde bola pozornosť venovaná štrukturálnym variáciám. Publikácia obsahuje pripojenú prílohu, v ktorej je okomentovaný a vysvetlený postup aplikovaných metód. Analýzy v tejto časti práce boli implementáciou týchto metód.

Dátami pre tieto analýzy bol výsledný súbor štúdie štrukturálnych variácií. Súbor bol vo formáte `vcf.gz` stiahnutý z databázy na *1000 Genomes FTP stránke*. `vcf` súbor bol kompresovaný vo formáte `gzip`, a je ho možné extrahovať použitím programov `WinRAR`, `7-Zip` alebo pomocou *online nástroja*. Vybranou metódou pre štatistickú analýzu je analýza hlavných komponentov (anglicky *principal component analysis PCA*), pomocou ktorej bolo možné študovať vzťahy medzi populáciami zahrnutými v tretej fáze 1000 Genomes Project. Dáta pre túto analýzu boli genotypy extrahované z `vcf`, a taktiež súbor obsahujúci zoznam vzoriek s určenou populáciou, kontinentálnou skupinou a pohlavím, ktorý bol stiahnutý z prílohy vyššie spomínanej publikácie. Na týchto dátach bolo prevedené PCA, ktoré rozdelilo Africké populácie od zvyšku a rozdelilo populácie z južnej Ázie, východnej Ázie a Európy do osobitných zhlukov, pričom Americké populácie prekrývajú Európske aj juho Ázijskej skupiny a teda nevytvárajú samostatný zhluk. PCA zobrazilo kontinentálnu populačnú štruktúru.

Populačná analýza dát z 1000 Genomes Project bola prevedená pomocou výpočtu heterozygoty a homozygoty u každého jedinca a výpočtom V_{st} pre každú štrukturálnu variantu. Heterozygotita a homozygotita boli stanovené pomocou genotypov delécií z `vcf` súboru. Heterozygotita/homozygotita bola odhadnutá ako počet heterozygotných/homozygotných udalostí u jedinca, ktoré sú definované genotypmi s rozdielnymi alelami. Táto analýza viedla k pochopeniu relatívnej diverzity populácií. Najviac heterozygotných delécií bolo zaznamenaných v Afrických populáciách, čo je úmerné zvýšenej rozmanitosti jednotlivcov z afrického kontinentu, a najmenej v populáciách z východnej Ázie. U homozygotných udalostí to bolo presne naopak. Na vyhodnotenie stratifikácie populácie sa použila štatistika V_{st} . Stratifikácia populácie umožňuje identifikovať adaptívnu selekciu. V_{st} porovnáva odchýlky vo frekvenciách alel medzi populáciami a umožňuje porovnanie viac alelických alebo multikópiových CNV (mCNV). Dáta použité na výpočet V_{st} boli genotypy delécií, duplikácií a mCNV (pomenované ako CNV v súbore VCF) získané zo súboru VCF, a súbor `sample.csv` ako v rámci vyššie uvedených metód na prístup k populačným kontinentálnym skupinám. V_{st} bolo určené pre každú variáciu medzi všetkými populáciami. Ak pre danú variáciu bola aspoň jedna hodnota $V_{st} \geq 0.2$, tak bola označená ako stratifikovaná. Najviac stratifikovaných štrukturálnych variant bolo identifikovaných u mCNV, následne u delécií a najmenej u duplikácií. Populačne stratifikované lokusy, ktoré doposiaľ neboli popísané sú cieľmi prípadného štúdia štrukturálnych variácií, ktoré podliehajú adaptívnemu výberu.

Posledná časť sa venuje bioinformatickej analýze, ktorá bola prevedená na ho-

mozygotných deléciách s cieľom nájdania exónových oblastí (CDS alebo UTR), ktoré boli úplne vymazané z dôsledku delécie. Zoznam použitých delécií bol stihaný z tabuľky **Table S6**, ktorá bola poskytnutá pri už spomínanej publikácii. CDS, UTR5 a UTR3 oblasti boli osobitne stiahnuté pomocou online nástroja UCSC Table Browser a následne boli programovo vyhodnotené kompletne vymazané regióny spolu s génmi, ktoré boli touto deléciou zasiahnuté. Program pracoval s RefSeq anotovanými génmi pre genóm GRCh37/hg19. Tie bolo následne potrebné preložiť do génových symbolov pomocou online nástroja DAVID. Program vyhodnotil 251 génov odpovedajúcich 215 deléciám. Pre tieto gény bolo následne stanovené RVIS skóre pre určenie, či boli homozygotne vymazané gény tolerantné mutáciám. To sa potvrdilo priemernou hodnotou RVIS, ktorá bola rovná 0.76, a ukázala, že geny sú tolerantné.

DECLARATION

I declare that I have written the Bachelor's Thesis titled "Bioinformatic analysis of single nucleotide polymorphisms in the 1000 Genomes Project database" independently, under the guidance of the advisor and using exclusively the technical references and other sources of information cited in the thesis and listed in the comprehensive bibliography at the end of the thesis.

As the author I furthermore declare that, with respect to the creation of this Bachelor's Thesis, I have not infringed any copyright or violated anyone's personal and/or ownership rights. In this context, I am fully aware of the consequences of breaking Regulation § 11 of the Copyright Act No. 121/2000 Coll. of the Czech Republic, as amended, and of any breach of rights related to intellectual property or introduced within amendments to relevant Acts such as the Intellectual Property Act or the Criminal Code, Act No. 40/2009 Coll., Section 2, Head VI, Part 4.

Brno

.....

author's signature

ACKNOWLEDGEMENT

I would like to thank my advisor, prof. Ing. Ivo Provazník, Ph.D., for consulting, patience and contributions which had led this thesis to its end.

Contents

Introduction	13
1 Basic of Genetics	14
1.1 DNA	14
1.2 Chromosomes	14
1.3 Genes	15
1.4 Human Genome	15
1.5 Genetic Variations	15
2 Variants Discovery	17
2.1 Sequencing and Reads Mapping	17
2.1.1 Sequencing Technologies	17
2.1.2 Reads mapping to the reference genome	18
2.2 Variants Genotyping	19
2.3 Variant Call Format	20
3 The International HapMap Project	22
3.1 Phase I	23
3.2 Phase II	24
3.3 Phase III	25
4 The 1000 Genomes Project	27
4.1 The Pilot	27
4.1.1 The 1000 Genomes pilot projects	27
4.2 Phase I	29
4.3 Phase II	30
4.4 Phase III	30
4.4.1 Structural Variant discovery	31
5 Databases	34
5.1 HapMap database	34
5.2 1000 Genomes Project database	35
6 Data retrieving and browsing Software	36
6.1 The 1000 Genomes Browser	36
7 Analysis of DNA Variation Data	39
7.1 Statistical Analysis	40
7.1.1 PCA	40

7.2	Population Analysis	43
7.2.1	Population Diversity	43
7.2.2	Vst analysis	45
7.3	Bioinformatic Analysis	49
	Conclusion	54
	Bibliography	55
A	1000 Genomes Project Populations	60
B	Downloading VCF file	61
C	Downloading CDS and UTR regions from UCSC Table Browser	62

List of Figures

2.1	VCF file exmaple	20
3.1	Haplotype	22
3.2	HapMap samples	23
4.1	Complex deletion groups	33
6.1	The 1000 Genomes Browser Page	36
7.1	SVs in VCF file	39
7.2	PCA	42
7.3	Compared analysis results with reference	45
7.4	Vst	47
7.5	Vst values compared	48
7.6	DAVID	50
7.7	Output of FindRegion Function	51
7.8	Deleted regions analysis	51
7.9	Regions	53
7.10	RVIS score	53
B.1	FTP page	61
C.1	UCSC Table Browser	62

List of Tables

1.1	Types of Genetic Variation	16
2.1	INFO column tags	21
4.1	Project results	29
5.1	HapMap database	34
5.2	1000 Genomes Project database FTP 1	35
5.3	1000 Genomes Project database FTP 2	35
7.1	Vst	47
7.2	Vst values compared	49
A.1	Populations	60

Introduction

The whole-genome sequencing made significant progress in bioinformatics, molecular biology and many other fields. Especially it has helped researchers to better understand common and rare diseases on a genomic level. It took several years and multiple scientific groups collaborating to get the final human genome sequence. The biggest success was made by inventing a sequencing platform named Illumina, which was able to sequence an entire genome in 50 hours, wherein the preparations took about 20 hours.

The beginning of the bachelor's thesis explains some basics of genetics, including variations. The next chapter is dedicated to variant discovery. Next two chapters are focused on the projects which are associated with the whole-genome sequencing and variant discovery, The HapMap Project and The 1000 Genomes Project respectively. The first two points of the bachelor's thesis assignment are mainly included in the Projects research as the whole genome mapping of DNA variations, and the genomic structural variants studied by allele frequencies and correlation patterns were performed by both of the Projects.

The HapMap was aimed to catalogue SNPs in the human genome to enable a better, faster and cheaper analysis of correlation patterns between variants and common or rare human diseases. The 1000 Genomes Project brought also other variant types as small deletions, duplications, structural variants and copy number variants known as CNPs.

Both projects have fully available databases on FTP Sites, which are describes in chapter 5. Further, the next chapter presents an online tool to search or download data produced by the projects.

The second part of the thesis studies the variant data produced by phase 3 of the 1000 Genomes Project. Statistical, population and bioinformatic analyses were run on data to show population stratification, gene-knockouts, relationships among populations or even the amounts of homozygous and heterozygous deletions within individuals.

1 Basic of Genetics

The first part is dedicated to fundamental questions about genetics and it is focused on Eukaryotic cell.

1.1 DNA

Deoxyribonucleic acid (DNA) is a molecule composed of four nucleotide bases: Adenine (A), Thymine (T), Guanine (G), Cytosine (C). Further is formed of deoxyribose sugar and phosphate. DNA is a double chain which is twisted into a double helix. These two chains, also known as 'coding' and 'template', are antiparallel and linked through nucleotides. One nucleotide base on one chain connects to a complementary base on the second chain, providing complementary base pairs (A come with T, and G with C). Thus the nucleotide sequence on one chain provides the nucleotide sequence of the second DNA chain. This complementary structure of DNA enables making its copies during DNA replication and cell division. Replication of DNA is formed according to the semiconservative model. Thus one half of a molecule DNA is preserved and the other one is synthesized. The length of DNA is given by base pairs and its denoted bp.[38]

1.2 Chromosomes

A human cell contains 23 pairs of chromosomes. Men have 22 pairs plus a chromosome XY. Women have also 22 but the 23rd chromosome is XX. XY and XX are named Gonosomes and they determine the sex of an individual. The rest of the chromosomes are named the Autosomes. Somatic cells have 2 lots of chromosomes, thus these cells are diploid. Gametes consist only from 1 lot of chromosomes and they are haploid. Fundamentally chromosomes are long threads of DNA (and protein) so they carry genetic information of a cell. They are located in a cell nucleus. A structure is best to be seen during the mitosis. In the centre is centromere which binds two pairs of arms. There are two types of arms, "p" and "q" ("p" for small, "q" for queue). Each of the chromosomes has an unique appearance. Their unchanging structural correspondence is frequently associated with a banding pattern (various sections of the chromosomes resemble as light and dark bands below specified laboratory conditions). These banding patterns have been used for a long time to determine a spot of particular genes. The location of a specific gene on the chromosome is defined as a locus. [38]

1.3 Genes

A gene contains information for a cell to read and use. The cell is capable to form a new protein because of a gene and its nucleotide sequence. Not every part is used in a process to create a protein. A gene can be divided into 3 main regions. The first region is at the beginning and its called a promoter. A promoter is an enzyme which leads to initiation of transcription of a specific gene by merging RNA (Ribonucleic acid) polymerase to this location. Transcription is the transmission of genetic information from DNA to mRNA. The third region is the opposite of a promoter because it is a terminator which is stopping the transcription. The most crucial region, the second region of a gene can be further divided into two regions. One of these regions is a coding one, therefore, it contains the genetic information to produce a protein. This region is named an Exon. The other, not coding region is called an Intron. According to [38], Intron is also helpful at transcription. It can link proteins that affect the regulation of the gene. The gene can assume several forms, called alleles. Each person carries one pair of the same gene, each inherited from one parent. Their interaction determines characteristics of an individual. Homozygosity refers to an identical allele of the gene on both chromosomes, and heterozygosity is when the gene occurs in two different alleles. A single gene can affect multiple characteristics at once and a single characteristic can be affected by multiple genes.[15]

1.4 Human Genome

The human Genome consists of a set of very long DNA molecules, each corresponding to one chromosome. There is around 25,000 genes within a molecule. The genome involves coding, and noncoding areas of DNA which do not encode any genes. The DNA sequence of the entire genome has been revealed in 2003 and has over three billion base pairs. The human genome is not the same for each individual and differs in populations. By knowing the differences, scientists are able to understand relations among subpopulations of humans, and it is also helpful to understand the origin of illnesses. In the past 20 years, the knowledge of the entire genome has helped in many fields as medicine, anthropology and forensics. [20]

1.5 Genetic Variations

Genetic variation also identified as genetic factor, is defined as a variation in the human genome which makes differences among individuals and their phenotypes. Each individual shares 99% of genetic information. Thus the difference between each individual depends on the remaining 1% of genetic variation. Variation is based on

different types of nucleotide bases on a specific location in a sequence. Any location in the genome where is more than one nucleotide base type within a population is called a mutation or polymorphism. Each nucleotide base variant is called an allele. Allele frequency describes the quantity of any allele which exists in human society. The allele frequency can be rare or common depending on the allele appearance among the population. Any individual has two lots of alleles froming homozygous or heterozygous genotype. Types of alleles in genotype are responsible for a phenotype. It is important to distinguish a mutation and polymorphism. If the variant is found in less then 1% of the human population it is a mutation and if the variant is more likely to be found (its appearance in population is more than 1%) then it is a polymorphism. Variations are studied as deviations from the reference genome. The reference genome is assembled from several individuals. Furthermore, table 1.1 shows and defines different types of variants. Single nucleotide polymorphism (SNP) is the most common variant (90% of all variants are SNPs). Another big section of variation studies is focused on copy number variant (CNV) also known as copy number polymorphism (CNP). CNV can be defined as a type of structural variability in form of deletions and duplications of genome segments that are not only numerous but also significant in their influence on organism physiology and function. CNV plays a role in predisposition to certain human diseases. At last, particular large variations have an impact on genome structure or function and they belong to a group of structural variations. [2] [18] [28] [38]

Tab. 1.1: Types and definitions of variations [2] [18] [33] [38]

Type	Definition
SNP	Single nucleotide polymorphism - base variation on a particular location in DNA sequences
CNV	Copy number variant - DNA segments are repeated and the number of repeats in the genome varies between individuals; typically deletions or duplications
INDEL	Insertion or deletion of a DNA sequence
Duplication	Duplication of DNA sequence (tandem and interspersed)
Inversion	DNA segment with a reverse base sequence.
Translocation	DNA sequence shift
MEI	Mobile element insertion - DNA sequence (transposon) which changes its location in genome
Microsatellite repeat	Short tandem repetitions - 1,2 or 4 repeating bases in a sequence which is shorter than 100 bp

2 Variants Discovery

The variant detection follows few steps: (1) sequencing, (2) sequence alignment to the reference genome, (3) quality control filters to exclude false-positive variants, (4) genotyping, (5) novel variants validation. The novel variants can be stored in the variant call format (VCF) which was developed for the 1000 Genomes Project. [12] [17]

2.1 Sequencing and Reads Mapping

This part is focused on sequencing methods used for DNA and human genome, and mapping reads which were produced by next-generation sequencing technologies. DNA Sequencing is the process of identifying the order of nucleotides in a DNA sample. Genome sequencing is more challenging and it takes much more time than a basic DNA sequencing. In order to detect variations between genomes, it is crucial to obtain parts of the sequences of each genome and align them to the reference genome. [10] [42]

2.1.1 Sequencing Technologies

Generations of Sequencing

First generation : *Sanger* [10]

Next-generation Sequencing (NGS) : *Illumina* is based on sequencing-by-synthesis chemistry. Firstly, adapters are added on both fragment ends, and one of these ends is attached on a flow cell coated with adapters and complementary adapters. Free end of a fragment is hybridising to the complementary adapter and creates a 'bridge'. The adapters are taking place of a primer used for PCR bridge amplification. Amplification is needed to obtain sufficient light signal to detect added bases by creating clusters of identical fragments. Each amplified fragment is sequenced with 4 reversible terminator nucleotides labeled with fluorescent dye representing specific base. Only one nucleotide is added during one cycle of synthesis because of its blocked site. The CCD camera identifies its base type via its fluorescent dye. The terminator group at 3' end of the nucleotide and the fluorescent dye are removed so the synthesis is repeated. [5]

Third generation Sequencing : *PacBio* or also called SMRT (Single-molecule real-time) sequencing provides real-time sequencing, much longer read lengths and faster runs than methods based on Sanger sequencing. The template used for PacBio is called SMRTbell and is made by ligating hairpin adaptors to the end of a double-stranded DNA molecule creating a closed circle. The entire process occurs in a

chip called SMRT cell, which has multiple zero-mode waveguide (ZMW) units with polymerase immobilized at their bottom and also with all four types of nucleotides labelled with a different fluorescent dye. The replication begins when the adaptor is bind to the polymerase, and as a base is held to a polymerase, the light pulse is created and can be detected in real-time to recognise a base type.[37]

Sequencing of human genome

High-throughput Sequencing Technology is characterized by low cost and high-throughput output. These technologies are mainly used for a whole-genome sequencing which enables a better comprehension of human diversity and disease. Some of the commercially available high-throughput sequencing platforms are technologies made by Illumina Company, Roche Company, ABI/SOLID Company, Helicos Heliscope and Pacific Biosciences. Illumina and PacBio were used during 1000 Genomes Project.[32] [35]

Sequencing methodes which are used for genome sequencing [31] :

Shotgun Sequencing is a modern method for sequencing. The analyzed DNA is randomly fragmented into short sections, which are cloned and then sequenced in both directions from the genome. The sections are overlapping thus the whole sequence can be assembled. For each segment, it is necessary to obtain several independent readings that defined the sequence coverage. Overlapping sequence segments are then combined to create the genome consensus. [9]

BAC Sequencing („Clone-by-clone“) splits the original DNA into overlapping fragments (about 150 kb) while the location of these fragments on the chromosomes is recorded to help with the sequence assembly. Then they are multiplied using a host bacterium (BAC-bacterial artificial chromosome) and again fragmented. A library is constructed from the multiplied fragments by sequencing. It has greater fault tolerance and is simpler in terms of calculation, but is slower and more expensive in terms of experiments.[43]

2.1.2 Reads mapping to the reference genome

The next-generation sequencing (NGS) technologies are producing millions of sequence reads which need to be mapped to the reference human genome to be used for genotyping and variant discovery. This process is crucial in creating biological context and meaning to the NGS data and to enable finding differences between the reference genome and unknown sequenced genome. One of the alignment algorithms is BLAST. BLAST shows similar sequences to one or multiple unknown

sequences on its input by comparing them to a huge database with annotated sequences. BLAST was mainly designed as a search tool but can be also used for finding diversity between two sequences.

The variant detection accuracy depends on how accurate is the NGS data alignment. A number of factors are affecting this process. First, all the adapters have to be extracted from the sequence reads. Second, alignment technologies can remove unreliable or nonhuman data within a read. This process is called 'hard clipping'. A 'soft clipping' is a process to improve indels detections and reduce false-positive SNP calls around indels. The ends of sequences have some amount of mismatched sequence. This sequence is not removed during this process, but it does not contribute to the alignment. Third, the read length is also one of the factors. Short reads can be mapped to multiple locations on a genome so there is a probability of reads being aligned at the wrong location. [8]

2.2 Variants Genotyping

Disease genetic studies aim to discover causative variants. Genotyping is a method created to identify a genotype, the genetic make-up of an individual, and can be used as an approach for such identification [27]. Genotyping is a huge advantage in the whole-genome association (WGA) studies, thus it helps to identify new genes involved in many common diseases. The high-throughput genotyping platforms are needed to assay for 1 million variants or more. The technologies which are suitable for WGA studies are Invader assays, The Perlegen Genotyping, Affymetrix GeneChips, and Illumina's Infinium Beadchips. GeneChip and Beadchip based platforms are highly successful in GWA studies and are describe below. [33]

Illumina's Infinium Beadchips

The microbead-base array (BeadChip) is made of an optical fiber bundle composed of around 50,000 individual fibers which are taking place of a substrate for microarray. Each fiber is etched to develop microwells to sustain a specific microbead that is covered with multiple copies of an oligonucleotide probe targeting a particular locus in the genome. [27]

There are two types of assays, Infinium I (10,000-100,000 assays) and Infinium II (up to 1 million assays). First steps are the same for both types starting with a whole-genome PCR amplification followed by hybridization to BeadChip of 50-bp-long capture probes. During the Infinium I assays, the probes contain locus-specific sequences with an allele-specific 3' terminal base. Genotype determination occurs through biotin-labelled nucleotides that are incorporated into an allele-specific

Variant call format (VCF) was developed for the 1000 Genomes Project and it is a generic format for storing variants with many annotations. Any variant type can be stored here. It enables to include millions of sites with genotypes data and annotations from numerous samples. VCF uses textual encoding with complimentary indexing. An example of a VCF format is in figure 2.1. VCF format is composed of a header segment and a data segment. The header contains meta-informations each starting with characters `##` explaining tags and annotations used in a file, and a TAB delimited field definition line beginning with the character `#` (`#CHROM` in figure 2.1(a)). This line names 8 obligatory columns which are CHROM, POS, ID, REF (reference), ALT (altenate), QUAL (quality), FILTER and INFO. These columns can be followed with FORMAT column and IDs of samples. FORMAT shows the order values within genotype table. For example `GT:GQ:DP` which stands for genotype, genotype quality and read depth. This is followed by a data segment showing values for each variant. VCF interpretation of variants is shown in figure 2.1 (b)-(f). The INFO column tags are explained below in the table 2.1. [12]

Tab. 2.1: INFO column tags explained [12]

DB	dbSNP membership
H3	association with HapMap 3
VALIDATED	validated by follow-up experiments
AN	total number of alleles in genotypes
AC	allele count for each ALT allele
SVTYPE	DEL for deletion, DUP dor duplication, INV for inversion etc.
END	position of variant ending
IMPRECISE	variant position is not known accurately
CIPOS/CIEND	confidence interval around POS and END positions for imprecise variants

3 The International HapMap Project

The Idea

Above all, the haplotype and linkage disequilibrium need to be explained. Set of alleles for different polymorphisms from the same chromosome are called haplotypes as it is shown in figure 3.1. Haplotype blocks can have a different set of alleles across individuals (SNPs). Linkage disequilibrium (LD) refers to nucleotides which travel together during chromosomal recombination (both are located in a single haplotype block). Alleles in a linkage disequilibrium travel together over generations so one allele provides information about the rest and can be used as a marker. LD can be measured between any two alleles on one chromosome using r^2 or D' measures. The HapMap project used the idea that some alleles within haplotypes are in complete LD with other alleles. Therefore one allele can represent multiple alleles that is in LD with and a haplotype can be identified by few SNPs called tag SNPs. [22] [38]

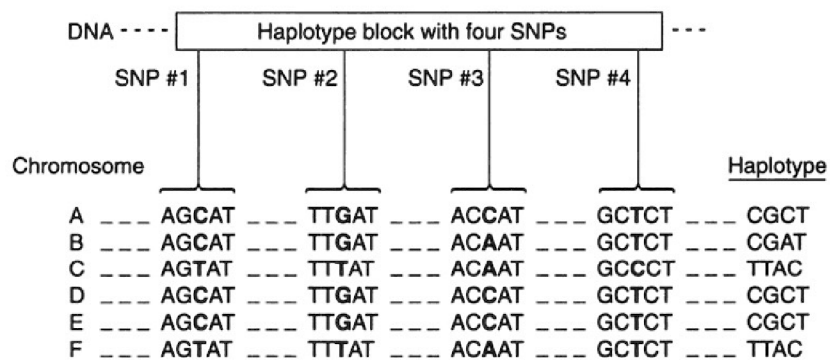


Fig. 3.1: A haplotype block with 4 SNPs. These 4 SNPs (boldly marked bases) create haplotypes for each chromosome **A-F**. [38]

Aim of the project

The project aimed to catalogue SNPs in the human genome and to enable discovery of causative variants causing diseases by making this data fully available for researchers. Scientists used the idea which has been already described above. By using tag SNPs and a linkage disequilibrium it was possible to genotype 200,000 - 1 million tag SNPs instead of genotyping over 10 million SNPs which are required to study the entire genome for correlation with a phenotype. This method enables researchers to study the entire human genome faster, cheaper and in a less complicated way. [22] [38]

Phase	ID	Place	Population
I/II	CEU		Utah residents with Northern and Western European ancestry from the CEPH collection
I/II	CHB		Han Chinese in Beijing, China
I/II	JPT		Japanese in Tokyo, Japan
I/II	YRI		Yoruba in Ibadan, Nigeria
III	ASW		African ancestry in the Southwest USA
III	CHD		Chinese in metropolitan Denver, CO, United States
III	GIH		Gujarati Indians in Houston, TX, United States
III	LWK		Luhya in Webuye, Kenya
III	MKK		Maasai in Kinyawa, Kenya
III	MXL		Mexican ancestry in Los Angeles, CA, United States
III	TSI		Toscani in Italia

Fig. 3.2: Samples used in three HapMap phases, including explanation of IDs used for this section. [25]

3.1 Phase I

The first, major phase of the project was to study genetic variations, find their location in a genome and compare their frequencies in different populations from around the world. The aim was to genotype one common SNP in the interval of 5 kb across the entire genome. The project analysed 269 DNA samples of unrelated individual and trios (parents and a child) from different populations presented in figure 3.2. Samples are described only with a population and sex identifiers without any other details about donors. 45 samples are enough to find 99% of haplotypes with a frequency above 5% in a population. LD studies can be done by using any individual samples. To evaluate a genotyping, ten 500 kb regions from ENCODE (Encyclopedia of DNA Elements) were sequenced. ENCODE contains common and rare SNPs which have been discovered and tested. [3] [22]

Genome-wide SNP discovery

The project produced a high-density map of SNPs. At the beginning of the project, there were around 2.6 million SNPs discovered and stored in dbSNP (The Public SNP database). To obtain more SNPs, the project applied shotgun sequencing from whole genome and whole chromosome (flow-sorted), expanded by a study of sequence traces produced by Applied Biosystems and data by Pelegen Sciences.

The number of SNPs in dbSNP expanded by adding 6.8 million SNPs, including 2.7 million SNPs whose each allele was seen at least 2 times ('double-hit' SNP). Since the length of haplotypes differs across the genome, it was necessary to use a hierarchical genotyping. First round of genotyping discovered around 600,000 SNPs each 5 kb long with a minor allele frequency (MAF) of at least 5%. New SNPs were tested on association with bordering SNPs by common LD measures D' and r^2 . Regions with a weak association (low or no LD) were genotyped again. More than 1 million SNPs were genotyped at the end. [3] [22]

Genotyping

Genotyping centres in 7 countries were using 5 high-throughput genotyping technologies: Third wave, Illumina, Perkin Elmer, Sequenom, Per Allele. Different types of technologies were used to compare the quality and cost of each sequencing. Genotype quality was tested during the entire process. For example, each centre received the same set of 1,500 SNPs for a genotyping test. Another example, the SNPs from each centre were shuffled and again genotyped by another centre. The quality control (QC) filters were also used and three cycles of quality assessment (QA) were applied to guarantee the reliability of genotyped data. [3] [22]

3.2 Phase II

The second phase has expanded the HapMap database by adding 2.1 million novel SNPs. 25-35% of all common SNPs ($MAF \geq 0.05$) were found. The SNP density was improved by finding one SNP per kb. [19]

Construction of the Phase II

Most of the novel SNPs were collected by utilising the Perlegen amplicon-based platform. Other technologies had taken place in the second phase as Affymetrix (GeneChip Mapping Array 500k, SNP array 6.0) and Illumina (HumanHap300, HumanHap550, HumanHap650Y). Reanalysing of SNPs from the first phase led to the discovery of 21,177 unreliable SNPs, which weren't included in the second phase. QC filters were used to eliminate further errors. A set of additional tests was applied on SNPs that passed the QC. Novel SNPs exhibited a lower MAF than in phase one which led to a rare variant discovery. A bigger number of SNPs had boosted a haplotype structure analyses, detection of recombinant hotspots, and especially it revealed earlier missed recombinant haplotypes. [19]

The use in Association Studies

The phase II showed an ability to discover more common SNPs and better selection of tag SNPs across the entire genome. By a simple pairwise tagging approach, it was discovered that twice more tag SNPs were needed to obtain all SNPs than in phase one, which discovered one-third as many SNPs. Perfect tagging ($r^2 = 1$) required 1.61 million tagSNPs in YRI population and around 1.05 million in CEU, CHB+JPT populations. Despite the high SNP density ($MAF \geq 0.2$), there were still some high-frequency SNPs which have been untagged. They could not be tagged because there is no SNP with $r^2 \geq 0.2$. These untaggable frequencies were mainly located within regions with a low LD and recombination hotspots. Recombination hotspots detection was also improved by 50% from a phase I. This enabled to explain the impact of genomic features on the distribution of recombination. [19]

3.3 Phase III

The HapMap 3 is another phase of the project. So far only common DNA variants ($MAF \geq 5$) were studied on a good level. There was a need to discover less common variants that have also a great impact on common disease studies. The HapMap 3 has studied common and rare variants besides of novel SNPs and CNPs (copy number polymorphisms) discovery. Using a genome-wide genotyping, 1.6 million SNPs were genotyped from 1,184 donors coming from 11 populations and 100 kb of regions were sequenced by PCR in 692 of these samples. 1,184 also includes samples from phase I and II, extra samples from CEU, CHB, JPT and YRI populations, and 7 novel populations (ASW, CHD, GIH, LWK, MKK, MXL, TSI [IDs explanation in figure 3.2]). Furthermore the project has studied population diversity among low-frequency variants. [4]

SNP Genotyping

The data were genotyped using Affymetrix 6.0 and Illumina 1.0 Million SNP mass arrays, following a quality control filters. [4]

Copy Number Polymorphism

CNP is a repeating section in genome and includes both duplications and deletions. 1,610 CNPs were discovered by genotyping and 856 passed QC filters. CNPs represented 3.5 Mb of sequence in each individual, and this covers 0.1% of the genome. More duplications were found than deletions. CNP distribution is at multi-kilobase scale, but most of them are found around 10 kb and with a low allele frequency

($MAF < 10\%$). Number of CNPs in each population is constant at common frequencies ($MAF > 10\%$) and vary below. [4]

The Population Differentiation

The population differentiation was measured by fixation index (Fst). The SNPs with Fst above 0.5 were determined as highly distinguished. The Fst value was tested between all populations and also between samples with the identical continental ancestries. Wright's approximate formula was used to calculate Fst:

$$Fst = (H_T - H_S)/H_T \quad (3.1)$$

where H_T is an assumed heterozygosity per locus of all populations, H_S is assumed heterozygosity of a particular population. Fst was counted only for SNPs with a $MAF > 5\%$. 28,215 common SNPs with a high population diversity ($Fst > 0.5$) were discovered for 49 various combinations of samples. These SNPs were divided into two categories, genic SNPs and nongenic SNPs. [4] [16]

4 The 1000 Genomes Project

The 1000 Genomes project started in 2007. The idea behind the project was a genome sequencing of 1000 individuals from different populations, and to understand the connection between the common diseases and genomic variations. The purpose was to develop a catalogue of genomic variations including variants with frequency $\leq 1\%$. The project was subdivided into 4 phases (a pilot, I, II, III). In 2015 the final phase of the project was published. Around 88 million variants were discovered, within 84.7 million SNPs, 2.6 million indels and 60,000 structural variants. All the population used during this Project are listed in appendix A. [14]

4.1 The Pilot

The purpose of the pilot phase of the project was to reveal and compare diverse methods for genome-wide sequencing through the high-throughput sequencing technologies. Three projects were realised: low-coverage (2-6 x) whole-genome sequencing of 179 individual from four populations, high-coverage (42 x) sequencing of two trios (mother-father-child), and exon-targeted sequencing of 697 individuals from 7 populations (coverage > 50 x). These project used samples from The International HapMap project. The goal was to define a site, allele frequency and local haplotype structure of 15 million SNPs, 1 million short INDELS and 20,000 structural variants. It was found that any individual had 250-300 variants that affect a normal gene function, and 50-100 variants that were already associated with inherited illness. Two trios project identified the frequency of de novo mutation across generations to be approximately 10^{-8} per bp per generation. [17]

4.1.1 The 1000 Genomes pilot projects

The low-coverage project used a whole-genome shotgun of a 79 unrelated individuals (59 from YRI, 60 from CEU, 30 from CHB and 30 from JPT). This project could identify shared variants on common haplotypes on a good level, but a low power to detect rare haplotypes and incorrect genotypes showed its disadvantages. The trios project also performed the whole-genome shotgun of two mather-father-child trios. Many types of variants in numerous genome regions were successfully identified. The exon project sequenced 8,140 targeted exons, from 906 random genes, in 697 individuals from 7 populations of African (YRI, LWK), European (CEU, TSI) and East Asian (CHB, JPT, CHD) ancestry, and provided reliable common, rare and low-frequency variant discovery but only in a targeted part of a genome. All three projects workflow procedures follow the same steps: (1) sequence alignment to the

reference genome, (2) quality control filters to remove false-positive variants, (3) genotyping, (4) novel variants validation through independent technologies followed by FDR (false discovery rate) determination. [17]

Variant Calling

All sequencing reads were aligned to the reference genome. To make sure all the reads have been placed within the correct regions where they originated from, most variant calling was limited to the accessible genome (a part of the reference after the incorrectly aligned regions are removed). The variant calls quality was evaluated firstly by using a base quality control, that rate the base diversity at any site. Secondly, at possible variant sites, the local realignment of all reads was done across all samples at once. The errors caused by misalignment were reduced. Thirdly, the data were analysed by comparing with other genotypes and variant calling algorithms, followed by forming a consensus. [17]

Variant novelty

Most of the novel variants were found in one panel (populations with the same ancestry). In contrast, the structural variants were mostly found in all panels. And most of novel variants were found in panels with African ancestry. The number of structural variants was decreasing with increasing variant length. Most of the novel SNP and structural variants were found between 10bp and 5 kb. Projects identified 50% of common short indels. The bigger number of samples and better coverage at the exon project enabled to discover variants with low frequencies ($MAF < 5\%$). [17]

Detection de novo mutations in trio samples

De novo germline mutations were discovered by sequencing (with a high coverage) of family members. The project was not strict with the FDR, but the second technologies are used to eliminate the false positive mutations afterwards. The trio project studied CEU and YRI trios. 3,236 and 2,750 samples were chosen where the disease is carried only by a child. The cell line DNA was re-sequenced for validation. The number of germline mutations decreased, 1,002 (CEU) and 669 (YRU) respectively. Another test for segregation to offsprings showed another number decrease, only 49 CEU and 35 YRI were positive for a germline mutation. Thus the mutation per bp per generation was estimated on 1.2×10^{-8} and 1×10^{-8} . The high number of false-positive mutation was caused by cell line or somatic mutations, so it was better to use DNA from primary tissue (blood) for large-scale studies. [17]

Rates of variant discovery

Tab. 4.1: Numbers of variants discovered across three Pilot projects [17]

	Coverage	Individuals	SNPs	Indels	Structural Variants
Low-coverage	3.6 x	179	14.4 million	1.3 million	20,000
Trio	42 x	6	5.9 million	650,000	14,000
Exon	50 x	697	12,758	96	

4.2 Phase I

Phase 1 performed a low coverage and an exome data analysis of 1,092 samples from 14 populations (Europe, East Asia, sub-Saharan Africa, America). This phase identified 38 million SNPs, 1.4 million short indels and over 14,000 larger deletions. 95% of common variants were already identified in the Pilot, but the low-frequency variants and the ones outside the exome weren't defined sufficiently. The low-frequency variants were important because they carry functional mutations, they are responsible for a genome sequences identification (for example they help distinguish shared variants from those private to families), and can explain strong geographic differentiation. [30]

Constructing an integrated map of variation

Genomes were analysed through low-coverage, exome and a high-density SNPs array data. In Pilot, low-coverage and exome projects showed a good power to identify variants except for the rarest SNPs and short indels. Thus the phase I ameliorated this research by using methods with a better quality variant calls and by blending SNPs, indels and large structural variants within a single work structure in 4 steps:

- (1) The purpose was to unite data from multiple centres and test them together. Primary data were integrated from low-coverage, exome and a high-density SNPs array data.
- (2) Data were aligned and read through multiple algorithms to get candidate variants. Each variant was then tested for its quality. The machine-learning method was used. The machine was learned through the high-density SNPs array data. This led to accurate rating of variants and FDR decrease.
- (3) Genotype probability served for the unification of evidence for each genotype on bi-allelic sites (0, 1 or 2 copies of the variant) in each sample.
- (4) Single genotype evidence was

usually weak while using the low-coverage data and can be very inconstant in the exome data. Therefore the statistical methods were used providing LD that enabled a haplotype (genotype) deduction. The phase 1 provided a high quality resource.[30]

Genetic Variation within and between populations

Populations are listed in appendix A annotated with III in phase column. Most of the common variants were described before (94%), but only 62% of low variants and 13% of rare variants were found previously. Phase 1 helped this discovery. According to the phase 1 results, almost all variants with a frequency above 10% were seen in all 14 populations, 17% of low-frequency variants were seen in populations with the same ancestry, and 53% of rare variants were seen within a single population. Some common variants differed strongly among populations with the same ancestry. African ancestry had three times more low-frequency variants as European and East Asian. Individuals across each population showed improvement in rare variants discovery. [30]

4.3 Phase II

Another 1700 samples were added during phase 2. The data was utilised for method improvement. First to improve the methods which already exist, second to create novel methods which enable research of other traits (multi-allelic variant sites, true unification of complex variation and structural variants). [1]

4.4 Phase III

The phase 3 of the 1000 Genomes Project studied 2,504 individuals from 26 populations (Africa (AFR), East Asia (EAS), Europe (EUR), South Asia (SAS), Americas (AMR)), including added samples of African (28% of novel variants) and South Asian (24% of novel variants) individuals. The methods developed during phase 2 were used to genotype and catalogue novel variants. Over 99% of SNPs with a frequency above 1% were discovered at the end. The project identified 88 million variants (84.7 SNPs, 3.6 million short indels, 60,000 structural variants) and phased them onto haplotypes. This enabled to move human genetic studies and common disease studies on a higher level. The samples were sequenced through a whole-genome sequencing, exon-targeted sequencing and high-density SNP microarrays genotyping of individuals with a first-degree relative where feasible. The machine-learning sorting of low-quality data was used to enhance the quality. Furthermore,

the project spreaded its focus on multi-allelic SNPs, indels and different types of structural variants. [6]

A typical Genome

A typical genome altered from the reference genome at 4.1 million to 5 million sites, and particularly 2,100 to 2,500 structural variants have modified almost 20 million bases. The populations with African ancestry exhibited most of the variants sites. Common variants were shared worldwide, rare ones were shared by closely related populations and 86% of variants were shared by populations from one continent. [6]

Putatively functional Variation

In general, there were 149 - 182 protein modifying variants, 10,000 - 12,000 peptide sequence changing variants, 459,000 - 565,000 variants overlapping regulatory regions. Variants influencing normal gene function were mostly seen in African populations. The European population had most of the variants per genome that were linked with a disease (around 2000) and rare diseases (24-30). [6]

4.4.1 Structural Variant discovery

Previous phases of the 1000 Genomes Project were focused on deletions while the phase 3 presented another type of structural variants (SVs). The SVs (including deletions, insertions, duplications and inversions) are causing most of the nucleotide changes and are connected with many human illnesses. SV genotyping was challenging but the improvement has been made and methods based on microarrays were replaced by short-read DNA sequencing with a bigger sample set. The samples were from unrelated individuals from 26 populations. The aim of this phase was genotyping of all main types of SVs (defined as DNA variants longer than 50 bp) across populations. 68,818 SVs were found and places into a map. [40]

Construction of phase 3 SV discovery

SV discovery of phase 3 SV release started by mapping Illumina WGS data from 2,504 individuals. Short reads with an average length of 100 bp were sequenced and mapped onto a reference genome (GRCh37) followed by genotyping using 9 different algorithms. The SVs accuracy and a false discovery rate (FDR) were estimated by using, for example, Affymetrix SNP6 and Illumina Omni 2.5 arrays both applying a long-read sequencing. A long-read sequencing call purifications enabled the additional 689 inversions and 9,132 small (<1 kbp) deletions to be included in calls, comparing with a SV set released by the 1000 Genomes Project marker

paper. The resulting SV callset includes 42,279 biallelic deletions, 6,025 biallelic duplications, 2,929 multiallelic copy-number variants (mCNVs), 786 inversions, 168 nuclear mitochondrial insertions (NUMTs), and 16,631 mobile element insertions (MEIs including insertions of Alu, L1 and SVA elements, 12,748, 3,048 and 835 respectively). Comparing with the Database of Genomic Variants (DGV), 60% of SVs were new, and 71% of SVs and 60% of collapsed copy-number variable regions (1 bp overlap) were new opposed to 1000 Genomes Project previous releases. The number increased mostly because of method improvements and a bigger number of populations. [40]

Population genetic properties of SVs

Populations genetic properties of SVs were examined by analysing all continental groups within this phase - Africa (AFR), Americas (AMR), East Asia (EAS), South Asia (SAS) and Europe (EUR). Variant allele frequency (VAF) investigations showed that 65% of SVs occurred at low-frequencies ($\text{VAF} < 0.2\%$) across all continental groups, while rare SV sites depended on each continental group. Almost all SVs were shared across all groups at $\text{VAF} \geq 0.2$. Considering variants being in linkage disequilibrium with nearby SNPs, 73% of SVs with $\text{VAF} > 1\%$ and 68% of rarer SVs ($\text{VAF} > 0.1\%$) have $r^2 > 0.6$, therefore they are in LD. By analysing deletion genotypes, it was shown that African ancestry exhibited 27% more heterozygous deletions, thus had the highest diversity among individuals than other continental groups, but had the lowest levels of homozygous deletions among all continental groups. East Asia exhibited the highest levels of homozygous deletions. PCA, which is a statical method, showed a continental population structure and admixture by analysing deletions within each population. Population stratification could be an option to identify adaptive selection, so SV differences in VAF amongst all populations were identified by Vst statistic which is highly correlated with Fst and quantify population differentiation. 1,434 highly stratified SVs ($\text{Vst} > 0.2$) were recognised, among which 578 intersected gene coding sequences (CDSs). [40]

Functional impact of SVs

A gene knockouts examination was performed on data of 5,819 homozygous deletions, and defined 240 genes (corresponding to 204 individual deletions) as 'dispensable' based on the analysis of homozygous losses in normal samples. More than 80% of these homozygous gene losses were novel comparing with previous studies. Furthermore, genes modified by homozygous loss were not very conserved and were tolerant to mutations. The functional impact of SVs was further quantified by expression

quantitative trait loci (eQTL) analysing 446 individuals from the GEUVADS consortium. [40]

SV complexity

The complexity of 29,954 deletions with determined breakpoints was also analysed in this release. It showed that 6% intersected another deletion with different breakpoints while 16% intersected another added sequence within the breakpoint interval. 1,651 deletions of 3,1 kbp with at least 10 bp long added sequence inside its limits were grouped into 5 classes shown in figure 4.1. 'MultiDup' refers to more apparent sequence duplicity within deletion boundaries and 'MultiDel' refers to 2 or more deletions that are divided with at least one sequence 'spacer' of up to 204 bp in length. [40]

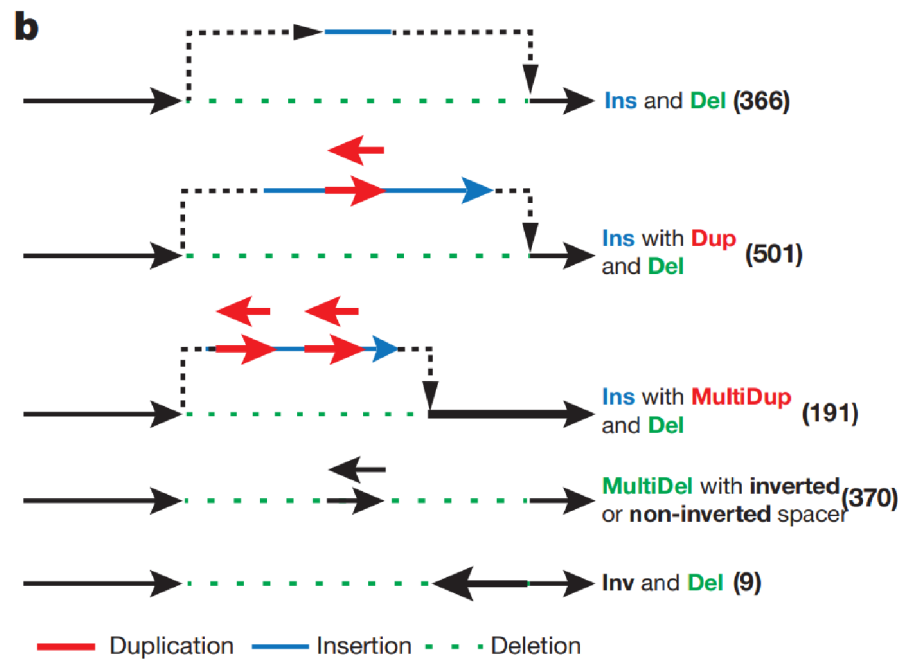


Fig. 4.1: Different complex deletion types. [40]

5 Databases

This chapter is focused on a structure of the International HapMap project and the 1000 Genomes project databases. Both databases are publicly available on NCBI (The National Center for Biotechnology Information) *FTP site*.

5.1 HapMap database

International HapMap Project *FTP* repository contains data released by the project and its particular phases. The database is split into multiple subdirectories which are listed and annotated in the table 5.1. [24]

Tab. 5.1: HapMap database subdirectories [24]

Subdirectory	Description
allocated_snps	SNPs mapped to the genome assembly stored individually for each chromosome
assays	information on assay used to genotype SNPs
cnv_data	data are coded with integer [0,1,2,3,4], 0-homozygous deletion, 1-heterozygous deletion, 2-duplicatons
frequencies	genotype and allele frequencies for SNPs, data files split in two, by center and by chromosome respectively.
hapmart	information about Hapmart tool for retrieving HapMap data
gbrowser	user manual for Gbrowser (browser is no longer available)
inferred_genotypes	genotypes inferred using in Silico methode
inter_chr_ld	interchromosome LD in HapMap release
jimwatsonsequence	SNP dataset
ld_data	datafiles with LD data compiled from genotypes data submitted by HapMap genotyping centers
mtDNA_and_chrY_haplogrups	samples
perlegen_amplicons	Perlegen amplicons summary
phase_3	genome-wide SNP genotyping in DNA samples from 11 populations and genotypes from 1115 samples
phasing	phasing information with data
presentations	HapMap tutorials
raw_data	raw data for HapMap project
recombination	data on recombination rates
sample_individuals	info on the family relationships in pedigrees for the samples genotyped in HapMap project
tmp	temporaly files
tscsnp	public data produced by the SNP Consortium project
xml_schema	defines elements, attributes and data types of xml data files

5.2 1000 Genomes Project database

The 1000 Genomes project has two mirrored FTP databases to store its data, *NCBI FTP Site* and *1000 Genomes FTP Site*. The structure of both sites is similar. There are 3 types of files at its top-level: README contains notes about site, CHANGELOG reports site changes over the years, and current.tree file shows all directories and files currently listed on site. Below these files, there are multiple directories listed in the table 5.2 for *NCBI FTP Site* and table 5.3 for *1000 Genomes FTP Site*.

Tab. 5.2: 1000 Genomes Project database directories on *NCBI FTP Site*

Directory	Description
alignment_indices	contains all previously produced alignment.index files
changelog_details	files with information about site changes
phase1	contains the analysis_results,data (samples) and technical subdirectories
phase3	contains samples in data subdirectory and data used in the SV discovery in integrated_sv_map subdirectory
pilot_data	data from the pilot
release	dated directories of analysis results sets with README files
technical	subdirectories for other data sets (simulations, method development files, reference genomes, etc.)
sequence_indices	contains all previously produced sequence. index files

Tab. 5.3: 1000 Genomes Project database directories on 1000 Genomes FTP *1000 Genomes FTP Site*

Directory	Description
changelog_details	files with information about site changes
data	all data moved to data_collection directory
data_collection	collection of data from different projects and the 1000 Genomes Project; each collection has README and index files (file/data descriptions) and data directory containing files organised by population and then samples
historical_data	created during site rearrangement to store files from the top level of the FTP site
phase_1	data from the phase 1
phase_3	data from the phase 3
pilot_data	data from the pilot
release	dated directories of analysis results sets with README files
technical	subdirectories for other data sets (simulations, method development files, reference genomes, etc.)

6 Data retrieving and browsing Software

This chapter will focus on working with the HapMap Project and The 1000 Genomes Project data. There are multiple online data browsers (Software) that enable both, access and analyse the data. As 99% of HapMap SNPs are found by the 1000 Genomes Project [7], it is possible to use one software that includes all data from the 1000 Genomes Project phase 3. [41]

6.1 The 1000 Genomes Browser

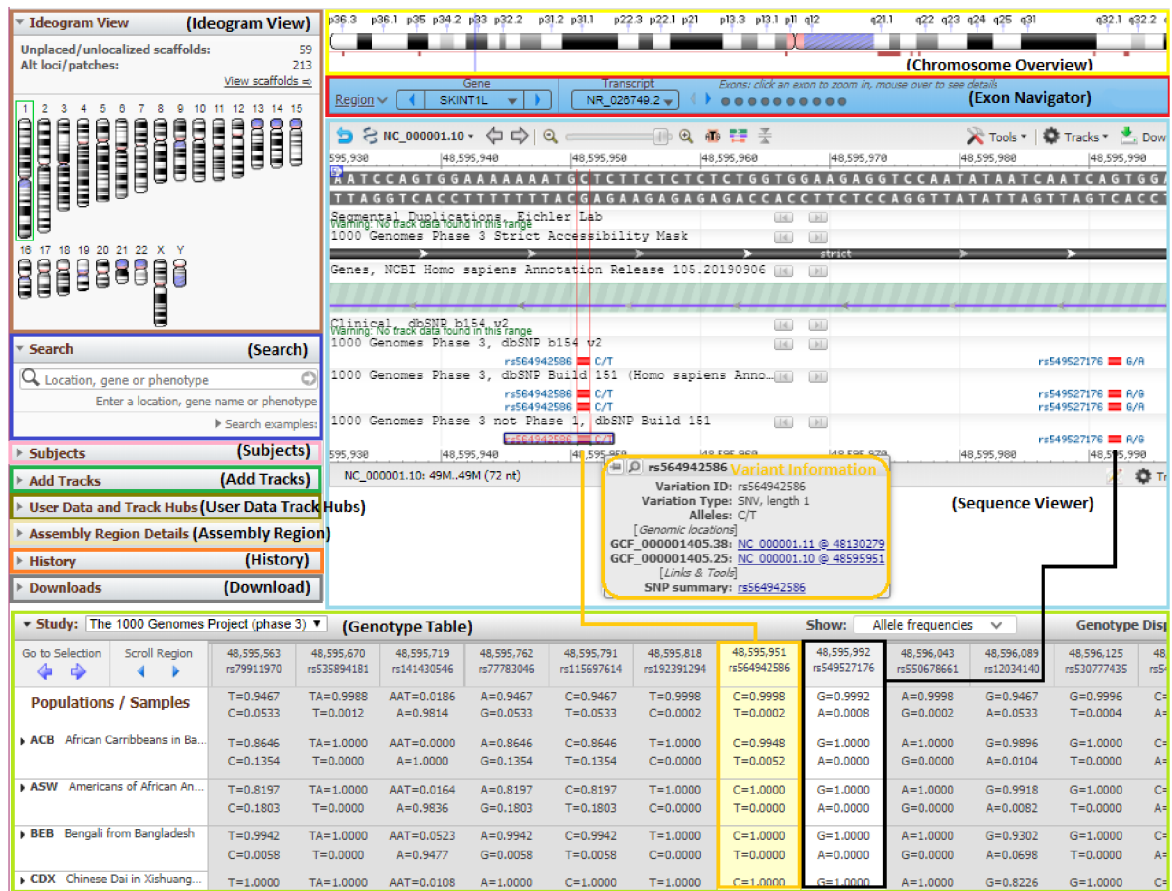


Fig. 6.1: A page overview with the various widgets highlighted in different colours.

The 1000 Genomes Browser is an online Software available on *NCBI Site* that allows researchers to explore variant calls, genotype calls and supporting sequence read alignments produced by the Phase 3 of the 1000 Genomes Project. The browser page is shown in figure 6.1. There are multiple widgets for data searching, analysing and downloading. [41]

Page widgets

Search widget is located in the left corner of the figure 6.1, and enables to find data by inputting a gene symbol, phenotype, variant ID from dbSNP and dbVar, cytogenic band and a diversity of formats for sequence coordinates. [41]

Ideogram View widget shows a chromosome which is currently analysed and it is placed below the Search widget. [41]

Chromosome Overview widget specifies a chromosome region which is shown in the Sequence Viewer. The blue box on a chromosome represents this region and it is possible to change its position and size. [41]

Exon Navigator widget is the blue bar placed under the Chromosome Overview widget. If the selected region contains multiple genes, it is possible to find and choose one of them in the menu labelled Gene. Double arrows on each corner can move the chromosome to the left or right and show previous or next gene in a region. After selecting a gene it is possible to also select an exon by clicking on one of the open circles or narrows. [41]

Sequence Viewer is a graphical representation of selected chromosome region. It is possible to reorder tracks or zoom in various ways and pan. Hovering over features in a track opens a menu with additional information. Tracks button placed in the right corner allows to add or remove tracks and upload personal data. [41]

Getotypes Table shows genotypes and allele frequencies across populations or samples within a population. It is possible to show allele frequencies, allele counts, sample frequencies or sample counts. Each column represents chromosome coordinates of variants within the selected region. Current displayed variants in the Sequence Viewer are in white columns and selected variant is in yellow. Hovering over the header row of any column brings up a tool tip with variant details. The second row shows global allele frequencies. Other rows represent specific population allele frequencies. Clicking on any population will roll up its samples. Next click on the specific sample within a population will show a box with BAM tracks which can be added to the Sequence Viewer. [41]

The Subjects widget can add tracks to the Sequence Viewer to see the read alignments formed for a specific sample. All available tracks are shown in a widget and can be filtered by using selected filters. [41]

Add Tracks supports GEO, SRR, or dbGaP additions to be used as an input. The addition will be placed to the Sequence Viewer if can be found in NCBI database. The track must correspond with a current track in a Sequence Viewer. [41]

User Data and Track Hubs widget allows to add personal data and to show them along the NCBI tracks in the Sequence Viewer, by uploading data or streaming

data from remotely-hosted data. [41]

Assembly Region Details shows another sequence path(s) which describes the selected region. The region has multiple sequence paths if its allelic diversity is complex. Clicking on any sequence in this widget will update the Sequence Viewer. Further, the widget has a GRC issues counter. The Genome Reference Consortium (GRC) is the project responsible for the amelioration of the human reference genome assembly. Tracks with these issues can be added to the Sequence Viewer. [41]

Download widget has two subsections. The first "SRA toolkit" provides downloading SRA (alignment) data for a region within the Sequence Viewer. There are also links to install the SRA toolkit, downloading parameters, and SRA Run Selector where the researcher can find all alignment data for selected samples. The second subsection provides downloading of genotype data (of a sample or aggregates by population) within a selected region in VCF format. The filter has two possibilities to filter the data, by either checked samples or population with checked samples. Genotype data can be downloaded via the Genotypes table widget. [41]

7 Analysis of DNA Variation Data

This chapter summarises the software analysis of variation data from the 1000 Genomes Project. The data and analysis methods which were used are from the *phase 3 Structural Variants release*. The release paper has a Supplementary Information file attached which includes explanations of used methods. Thus all the analyses from this chapter are basically an implementation of the 1000 Genomes Project studies. The implementation was done in a programming language named R, and the structural variant studies were presented by bioinformatic, statistical and population analyses.

The structural variant data were downloaded from the *FTP site* in VCF format and imported into environment with `vcfR` package. Precisely, the file containing all structural variants generated by the 1000 Genomes Project was downloaded in April 2020. Variants IDs are within `ID.xlsx` file. The data are not included in attached files because of its size, so all steps of downloading this dataset are explained in appendix B. In addition, a file `samples.csv` containing information (ID, continental group, population and gender) about all samples from this phase of the 1000 Genomes Project was downloaded from the *SV release paper* supplementary material named `Table S1`. The data from this file were mainly used to divide samples either by populations or continental groups. All populations with annotated continental group are listed in appendix A.

The uploaded VCF file data are shown in figure 7.1. The structural variant data from the VCF file were further examined by `vcfR` package functions to obtain variant IDs, genotypes or other features which were included in the INFO column of the VCF file (SV type, allele frequencies, allele count).

```

      CHROM POS      ID      REF ALT      QUAL FILTER
a [1,] "1"    "645710" "ALU_uary_ALU_2" "A" "<INS:ME:ALU>" NA    NA
  [2,] "1"    "668630" "DUP_de1ly_DUP20532" "G" "<CN2>" NA    "PASS"
  [3,] "1"    "713044" "DUP_gs_CNV_1_713044_755966" "C" "<CN0>,<CN2>" NA    "PASS"
  [4,] "1"    "738570" "UW_VH_21763" "G" "<CN0>" "100" "PASS"
  [5,] "1"    "766600" "UW_VH_5595" "G" "<CN0>" "100" "PASS"
      FORMAT HG00096 HG00097 HG00099 HG00100
  [1,] "GT" "0|0" "0|0" "0|0" "0|0"
  [2,] "GT" "0|0" "0|0" "0|0" "0|0"
  [3,] "GT" "0|0" "0|0" "0|0" "0|0"
  [4,] "GT" "0|0" "0|0" "0|0" "0|0"
  b [5,] "GT" "0|0" "0|0" "0|0" "0|0"

```

Fig. 7.1: **a** Example of structural variants stored within VCF file. Each row represents a single structural variant with its features in separate columns. **b** An example of SV genotypes stored in VCF file.

7.1 Statistical Analysis

Principal component analysis was done to explore relationships among different populations which were used in phase 3 of the 1000 Genomes Project. It is necessary to understand the main idea of using PCA in genetic studies before the analysis method is explained.

7.1.1 PCA

Principal component analysis (PCA) is a statistical method for investigating datasets, which have a great number of measurements, by reducing them into only a few principal components (PCs) that show the main patterns of the dataset. The first PC is the mathematical combination of measurements with the biggest variability in the data. PCA can be further used in genetic studies. It was primarily used for human population migration analysis but has also shown a good potential to explore differences in ancestry among populations and samples by analysing variants. [36]

Basic PCA computation

The method description is taken from [39] and [26]. The first step is data normalization by subtracting the mean \bar{X} (average of each dimension) from each value X_i within a dimension j as shown in equation 7.1.

$$X_{ij}^* = X_{ij} - \bar{X}_j \quad (7.1)$$

Step 2 is a calculation of the covariance matrix C using the formula:

$$C = \frac{X^{*'} X^*}{n - 1} \quad (7.2)$$

where C is a matrix with n rows and n columns, and n is number of dimensions.

Next step is the computation of the eigenvectors and eigenvalues of the covariance matrix to identify principal components. The relation between covariance matrix C , eigenvectors a and eigenvalues λ :

$$Ca - \lambda a = 0 \Leftrightarrow Ca = \lambda a \quad (7.3)$$

The eigenvectors are then ordered by their eigenvalues in descending order, so this step enables to eliminate some components depending on their variability. Less significant components have lower eigenvalues.

The last step is to transform the initial data among the principal component vectors by multiplying them by eigenvectors.

Method

The PCA was used to show differences between all 26 populations from phase 3 (appendix A). The R script `pca.R` is formed of one main function and one subfunction. `PCA` function has on its input VCF file and a `samples.csv` file, the output is a list class `prcomp` which contains the `sdev` (standard deviations of the PCs), `rotation` (matrix whose columns contain the eigenvectors) and `x` (the samples coordinates in PCs dimension).

The data used for PCA are deletion genotypes. Thus the `PCA` function extracts deletion genotypes for each individual (2,504) and transforms them into 0,1,2 coding by calling function `transform` (homozygous reference \rightarrow 0, heterozygous \rightarrow 1, homozygous alternate \rightarrow 2). Furthermore, genotypes were normalized by Patterson. Briefly, the mean value μ was estimated for each deletion and genotypes were centred around μ :

$$x_{ij} = x_{ij} - \mu_i \quad (7.4)$$

where i refers to a particular deletion and j is a sample.

Genotypes were then divided by $\sqrt{p(j)(1-p(j))}$ where $p(j)$ is an estimate of allele frequency of a deletion site and equals $\frac{c}{n}$. [40]

After the data were centered and normalized, the PCA was performed on entire set of populations and computed by `prcomp` function from `stats` package, which was the only function that could have handled the size of used dataset. First two columns from `x`, that is one of 3 `prcomp` function outputs, were plotted using `ggplot2` package and represent first two principal components PC1 and PC2.

Results

The result of PCA is shown in figure 7.2 [top]. First and second principal components reflect population stratification among entire set of samples. Each population is plotted with different point shape or color, while populations from the same continental group are plotted with the same color. From the figure, it is obvious that PCA separated African populations from non-African along PC1 and showed an admixture of these non-African populations. Further, it is possible to see that PC2 separated European, South Asian and East Asian populations into individual clusters while the American populations overlap European and South Asian populations. Furthermore, PC2 also separated East Asian populations from South Asian, European and American. In summary, PCA has showed the continental population structure and admixture.

Comparison with the 1000 Genomes Project

Analysis output was compared to the reference obtained from the 1000 Genomes Project. The population clusters are correct but the axis span is larger. It could be caused by missing some step in the calculation caused by not being annotated in the supplementary file of *the phase 3 release* or by using a wrong function to compute PCA in R. But the used function `prcomp` was giving the most similar result to reference. Furthermore, there are also two outliers from CHB population (dark blue) in analysis result. It could be caused by not filtering them or basically they were cut off from the plotting figure in the reference result.

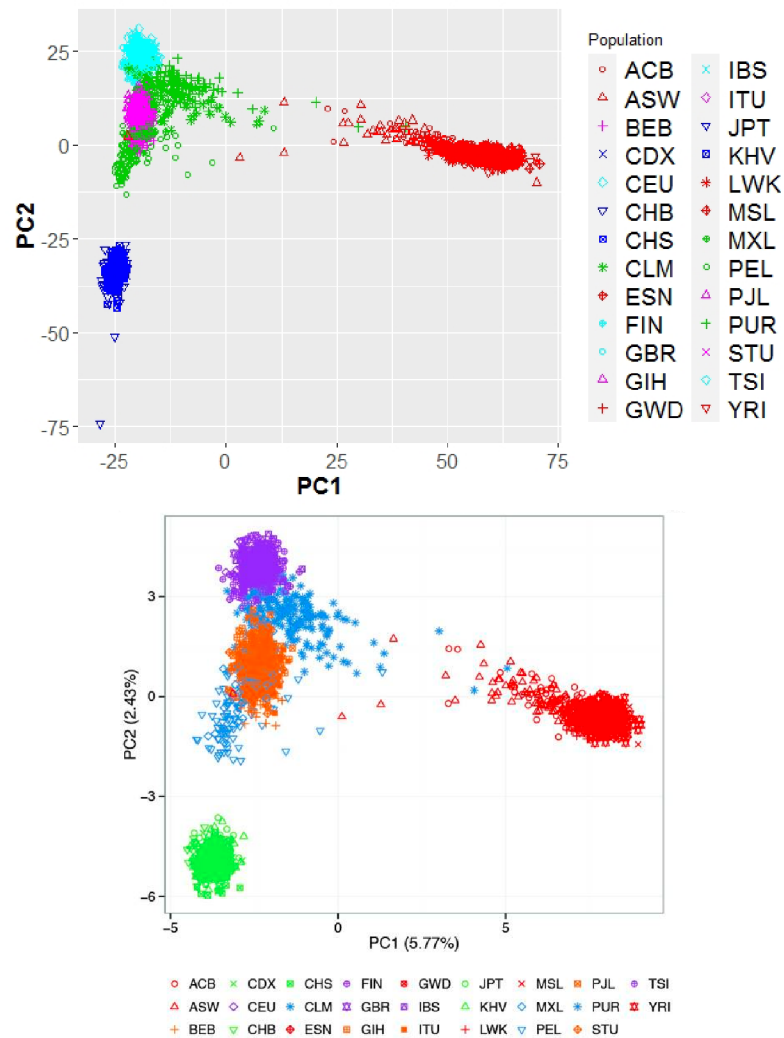


Fig. 7.2: Principal component analysis plot of principal component 1 and 2 for deletions. Analysis output (top) and reference output (bottom). [40]

7.2 Population Analysis

The purpose of population genetic analysis is to explore for example the diversity of a population or its stratification on a given dataset of samples from different continental groups or with different ancestries. The differences between any populations are caused by the existence of multiple alleles at different gene loci. [23]

The population analyses on a dataset of populations examined by phase 3 of the 1000 Genomes Project were done by calculating heterozygosity and homozygosity within each population, and by measuring Vst for each structural variant.

7.2.1 Population Diversity

A relative diversity of all particular populations was estimated by per-individual SV-homozygosity and SV-heterozygosity calculations for deletions. SV-homozygosity is determined as a number of homozygous events within an individual, and SV-heterozygosity as a number of heterozygous events.

Homozygous and Heterozygous

Each chromosome within a human body has another copy, thus there are two copies of each gene located on a locus. These two genes are called alleles. The person is homozygous for that trait if the alleles are the same (AA or aa), while if they are different (Aa or aA), the person is heterozygous. [29]

Method

The script `HomoHeterozygosity.R` for the homozygosity and heterozygosity computation is constituted from a single function `HomoHeterozygosity` with a VCF file on its input. The function has also used data from the `samples.csv` file. Data for the analysis were extracted deletion genotypes from the VCF file. Samples used for this study are listed in appendix A. Heterozygosity and homozygosity were counted for each individual separately. As shown in figure 7.1, the genotypes are coded as 0|0, thus the homozygous genotypes were defined as 1|1, 2|2, 3|3. 0|0 indicates that the deletion site is not located within a particular individual. Genotypes with different alleles as 0|1, 0|2, 0|3, 0|4, 1|0, 1|2, 2|0, 2|1, 3|0 and 3|2 were defined as heterozygous events. So the heterozygosity/homozygosity within an individual is estimated as a number of the total heterozygous/homozygous events. The function `HomoHeterozygosity` creates two separate data frames for homozygosity and heterozygosity counts. Each population is within a column and plotted in complex figures that show the number of homozygous deletions in each population, heterozygous respectively.

Results

A figure 7.3 [top] displays results of the computation of homozygous and heterozygous deletion within each individual in this study. The top figure belongs to homozygous events, and the bottom figure represents heterozygous events. On average, the highest levels of SV-homozygosity were exhibited by East Asian populations with around 430 homozygous events and the lowest by African populations with only 230 events. Apart, the maximum homozygosity is displayed by Mexican Ancestry in Los Angeles, California (MXL) followed by Peruvian in Lima, Peru (PEL), their homozygosity is above 450 and they are both from American continental group. Other populations from American group didn't show increased homozygosity numbers, and their values are similar to South Asian and European populations.

In contrast, the highest levels of heterozygous events were shown by the African population with two increased values within Esan in Nigeria (ENS) and Luhya in Webuye, Kenya (LWK) populations whose maximum is around 2,000, commensurate with the increased diversity of individuals from the African continent. Another excess in the heterozygous events can be seen within the population from Toscani in Italy (TSI) with almost 1,750 events. East Asian populations present the lowest number of heterozygous events, around 740 precisely. South Asian and European populations have similar number of heterozygous events. American population has after African population highest levels of heterozygosity.

Comparison with the 1000 Genomes Project results

To determine analysis accuracy, the data were empirically compared with results obtained by the 1000 Genomes Project in figure 7.3. Both homozygous and heterozygous deletions figures are very similar to the reference except outliers in some of the populations. The difference between the numbers of homozygous events in analysis figure 7.3 [top] and reference figure 7.3 [bottom] is in their maximums. The population with maximum homozygous events is MXL (America) followed by PEL (America) in analysis figure, while in reference is the PEL population first. The heterozygous events along populations have more outliers in the analysis figure compared to the reference. The population with the most numerous heterozygous events in reference is ASW (Africa) while in analysis figure it is LWK (Africa) followed by ESN (Africa). Other excesses can be seen within TSI (Europe), FIN (Europe), CLM (America) and YRI (Africa). The outliers could be caused by not filtering numbers of events or by using a different dataset than the 1000 Genomes Project have used. Although the analysis has followed the method steps which are annotated in the supplementary file attached to *the phase 3 SV release paper*.

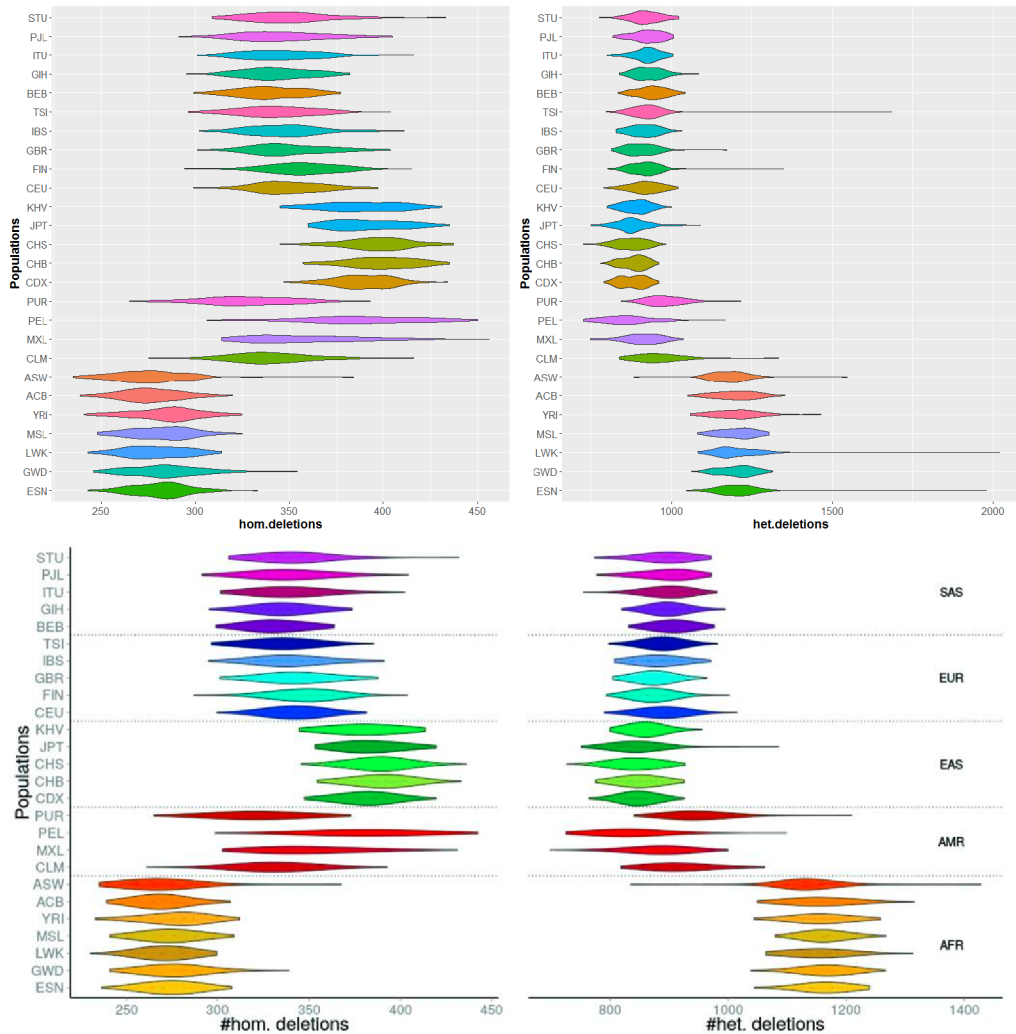


Fig. 7.3: The analysis output (top) and the reference output from the 1000 Genomes Project (bottom). Both figures show the number of homozygous (left) and heterozygous (right) deletions within each population. [40]

7.2.2 Vst analysis

Vst statistic was used to assess population stratification. Population stratification enables to identify adaptive selection. Thus a population stratification can be indicative of loci under adaptive selection. Vst is highly correlated with Fst. Vst and Fst statistics compare the variance in allele frequencies between populations with Vst allowing comparison of multi-allelic or multicopy CNVs. Values are between 0 and 1. Higher Vst values suggest differentiation between populations, lower values indicate that populations are very similar. Undescribed population-stratified sites are potential objectives for the eventual examination of SVs undergoing adaptive selection or genetic drift. [40] [44]

Method

Data used for Vst computation were deletion, duplication and mCNV (named as CNV in VCF file) genotypes obtained from VCF file, and `samples.csv` file as within the methods above to access population continental groups of the dataset (appendix A). Vst was calculated through several functions located within a script named `vst.R`. Vst was calculated for deletions, duplications and mCNV individually by estimating its value for a specific variant site between all continental groups (AFR, AMR, EUR, SAS, EAS). Thus each variant site had 10 Vst values among pairwise continental groups.

`Vst_df` function had on its input all three calculated Vst data frames for deletions, duplications and mCNV separately and created a new data frame with total numbers of deletions, duplication and mCNV sites and a number of variants which had at least one $Vst \geq 0.2$. This data frame was used for results plotting.

The individual data frames of deletion, duplication and mCNV Vst values were obtained with `var_type_vst` function. This function has on its input single variant type (DEL/DUP/CNV) and calculates Vst value for each variant of the inputted type between all continental groups.

Vst calculation was done by using function `Vst`. Inputs were 2 continental groups and a variant type. Genotypes for the particular variant type were extracted from the VCF genotype table. It was necessary to transform genotype coding as shown in figure 7.1 to 0, 1, 2 coding by using function `transform` (homozygous reference \rightarrow 0, heterozygous \rightarrow 1, homozygous alternate \rightarrow 2). The Vst value was then calculated for each variant site as follows:

There are two populations A and B.

Before Vst was computed it was necessary to estimate population variance σ_x^2 for population A and B using the following formula [13]

$$\sigma_x^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu_x)^2 \quad (7.5)$$

where σ_x^2 represents particular population (A or B), N is a number of individuals in population, x_i are samples and μ_x is the mean of the vector with samples for a particular variant site.

Next, the total variance σ_T^2 was calculated by binding populations A and B into one, so for a specific variant site, samples from populations A and B were putted into one vector that was used for σ_T^2 computation of population variance using the equation 7.5.

Vst was calculated as

$$V_{st} = (\sigma_T^2 - V_S) / \sigma_T^2 \quad (7.6)$$

where σ_T^2 is a total variance and V_S is the population-specific variance calculated using equation

$$V_S = \frac{(\sigma_1^2 * N_1 + \sigma_2^2 * N_2)}{(N_1 + N_2)} \quad (7.7)$$

where σ_x is a population variance of population x, N_x is a number of individuals within population x. [44]

Results

For each variant site, the V_{st} was estimated among all pairs of continental population groups. Furthermore, obtained V_{st} were analysed for high stratification. The variant site is stratified if at least one $V_{st} \geq 0.2$, thus if at least one population pair satisfies this condition, it indicates population stratification and possibility of SV going under adaptive selection.

A table 7.1 and figure 7.4 presents numbers of high- V_{st} events identified in each class. The highest percentage of stratified SVs out of total SV number was exhibited by mCNVs, followed by DEL, 3.5% and 3.2% respectively. Only 0.2% of DUP were stratified.

Tab. 7.1: Variant stratification.

	DEL	DUP	mCNV
total number of SV	42279	6025	2929
$V_{st} \geq 0.2$	1342	12	101

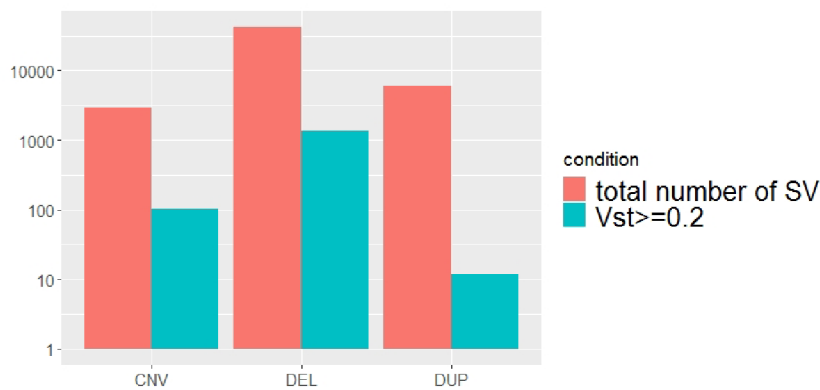


Fig. 7.4: Number of $V_{st} \geq 0.2$ compared to total number of variants separated by its type. Pink boxes represents total number of variants and blue high stratification variants. Y-axis is log-scaled.

Comparison with the 1000 Genomes Project results

The results were compared with the reference results from the 1000 Genomes Project to determine their accuracy. The reference results were downloaded from *the phase 3 SV release paper* supplementary files (Table_S5). The compared results in figure 7.2 show that only duplications were counted correctly, there is an excess of 30 sites within the deletion analysis result and a lack of 10 sites within mCNV analysis results. These values are also shown in figure 7.8 [top]. The reference source has analysed 42,441 deletions, 6,120 duplications and 2994 mCNVs instead of 42,279, 6,025 and 2,929 stored in VCF file and used for the method implementation. The figure 7.5 [bottom] interprets the result differences in percentage, considering the different quantity of SV sites, where the maximum difference between analysis and reference results is estimated at 0.2%. Thus the differences could be caused by analysing a different set of SVs.

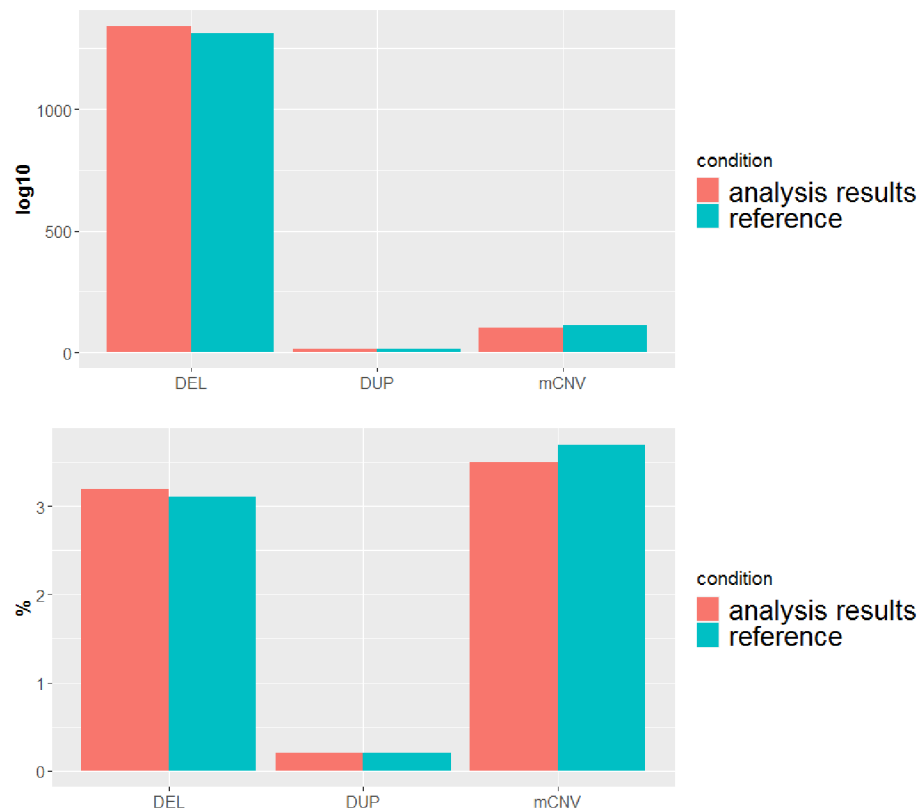


Fig. 7.5: Top graph compares portions of stratified sites discovered by analysis and the reference. Bottom graph shows percentage of discovered stratified sites within analysis and reference datasets.[40]

Tab. 7.2: Numbers of stratified sites discovered by analysis and the 1000 Genomes Project. [40]

	DEL	DUP	mCNV
analysis results	1342	12	101
reference	1312	12	111

7.3 Bioinformatic Analysis

The bioinformatic analysis was run on homozygous deletion to discover gene knock-outs occurring naturally in populations. After the genes were found, the RVIS score was computed for each deletion site that has completely deleted at least one exon, thus affected a gene. These identified genes seemed to be 'dispensable' on the basis of the observatory of homozygous losses in normal individuals.

RVIS score

RVIS (Residual Variation Intolerance Score) is a gene score used for human sequence data interpretation. The score ranks genes by whether they have more or less common functional genetic variation relative to the genome-wide expectation given the amount of apparently neutral variation the gene has. Genes with a positive RVIS score are highly tolerant of mutation while the gene with a negative score has less common functional variation and are marked as 'intolerant'. For example, the RVIS score for ATP1A3 shows a negative value -1.53 and a percentile of 3.37% showing that it belongs to 3.37% of the most intolerant human genes. The RVIS scores can be found and downloaded from the *Genic Intolerance site* for any gene. [21]

Data

Analyzed 5,819 homozygous deletions were downloaded and extracted from a supplementary table (**Table_S6**) of phase 3 SV release *paper*. First three columns of this table stored chromosome number, start and end position of each deletion site.

To find out if the deletion site has completely removed an exon (untranslated region (UTR) or coding sequence (CDS)) it was necessary to download CDS and UTR positions for an entire genome region. The data was downloaded from the Table Browser hosted on the *UCSC site*. The 1000 Genomes Project used RefSeq gene annotations for GRCh37/hg19, so it was necessary to change browser criteria. All steps explaining how to download regions from UCSC Table Browser are in appendix C. The output file format was set to BED so it could be easily uploaded

to R environment. The BED file was constituted with chromosome name, start and end positions, and annotated genes for particular regions.

The gene names in BED file were annotated with RefSeq IDs. *DAVID site* was used to transform gene IDs into gene symbol for better interpretation. DAVID is an online tool and to get gene symbols for a list of RefSeq genes these steps were followed: Gene names were extracted from BED files and saved as a csv table in R (`writexl` package) and uploaded into DAVID. Before a file was submitted, the identifier was set to REFSEQ_RNA and list type to Gene List. Next step was to choose with which DAVID tool a list should be analyzed. The gene ID conversion tool was chosen and OFFICIAL_GENE_SYMBOL was set as a conversion type. The output is shown in figure 7.6, where the first column is an old RefSeq ID and the second column is assigned gene symbol. The file was downloaded and saved as `gene_SYMBOL.bed`.

Gene Accession Conversion Tool [Help](#)

Gene Accession Conversion Statistics [Download File](#)

Conversion Summary			Submit Converted List to DAVID as a Gene List		Submit Converted List to DAVID as a Background	
ID Count	In DAVID DB	Conversion	From	To	Species	David Gene Name
44338	Yes	Successful	NM_018002	OXR1	Homo sapiens	oxidation resistance 1(OXR1)
0	Yes	None	NM_018000	MREG	Homo sapiens	melanoregulin(MREG)
0	No	None	NM_018006	TRMU	Homo sapiens	tRNA 5-methylaminomethyl-2-thiouridylate methyltransferase(TRMU)
0	Ambiguous	Pending	NM_004211	SLC6A5	Homo sapiens	solute carrier family 6 member 5(SLC6A5)
Total Unique User IDs: 44338			NM_004212	SLC28A2	Homo sapiens	solute carrier family 28 member 2(SLC28A2)
Summary of Ambiguous Gene IDs			NM_018004	TMEM45A	Homo sapiens	transmembrane protein 45A(TMEM45A)
ID Count	Possible Source	Convert All	NM_018003	UACA	Homo sapiens	uveal autoantigen with coiled-coil domains and ankyrin repeats(UACA)
All Possible Sources For Ambiguous IDs						
Ambiguous ID	Possibility	Convert				

Fig. 7.6: Output page of DAVID's ID conversion tool. In table on the right, the first column is inputted data and the second are names transformed into gene symbol format.

Method

The program in `gene.R` script has two subfunctions which are called by the main function `findRegions`. Main function inputs are homozygous deletion regions and names of CDS, UTR5 and UTR3 regions files within a vector `UCSC_outputs`. CDS and UTR regions files are not included in the attached files because of their size, so the downloading manual is in appendix C. In attached files, there are test data containing regions only from chromosome 1, so the program can be run either with these data or an entire dataset with downloaded CDS and UTR regions. Deletion regions that have removed at least one exon are found by analysing if the CDS/UTR5/UTR3 regions are within a deletion interval with `findOverlaps` function

(IRanges package). If deletion overlaps CDS or any of UTR regions, the region type is saved into a region column in an output data frame shown in figure 7.7. Further, any gene located in this region interval is extracted from the CDS/UTR BED files and transform by the function `RefSeq` from RefSeq ID into gene symbol using DAVID's outputted file. Furthermore, the function `RVIS` estimates a minimum RVIS score of all affected genes per a particular deletion region. RVIS score is downloaded from the *Genic Intolerance site* and stored in `RVIS.bed` file.

45	chr1	16809854	16811588	UTR3	CROCCP3	NA
46	chr1	17676165	17677662			NA
47	chr1	19599364	19614236	CDS,UTR3	AKR7A3,AKR7L	97
48	chr1	19612424	19614161	CDS	AKR7A3	97
49	chr1	19647187	19648342			NA

Fig. 7.7: `FindRegions` function output example. Columns: chromosome number, start position, end position, deleted region, gene, RVIS score

Results

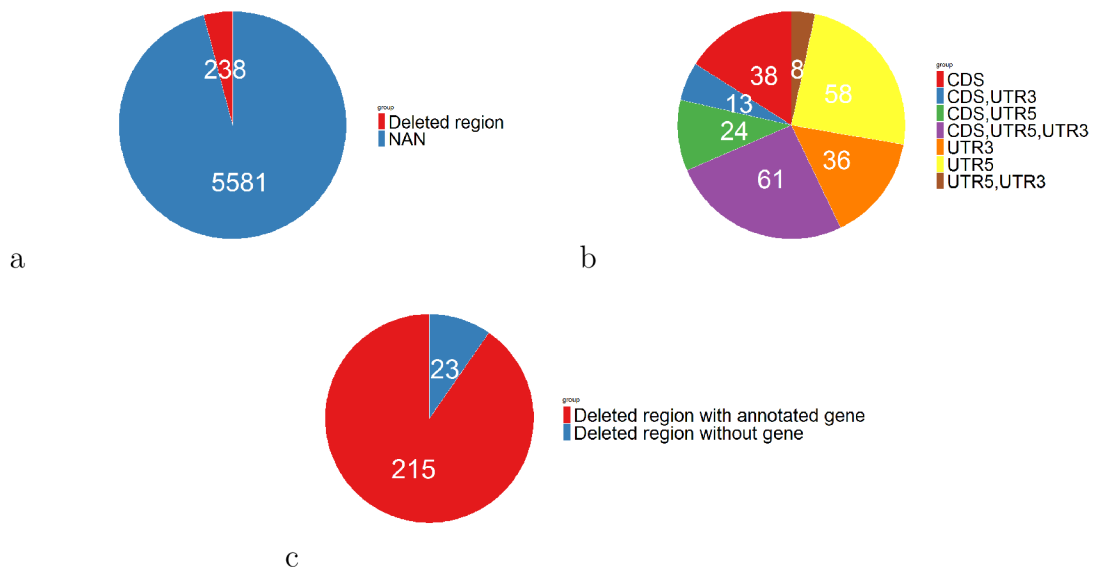


Fig. 7.8: Deleted regions analysis. **a** number of deletions sites that deleted CDS/UTR regions compared with number of sites where no region was removed. **b** numbers of deleted region types annotated at deletion sites. **c** portions of deleting region sites with annotated gene and without.

All result are within a single data frame. The program was able to find 238 deletion sites that deleted at least one exon as it is shown in figure 7.8 (a). The

region affecting deletions were only 4% of all 5,819 homozygous deletions. Although, 23 out of 238 did not have an annotated gene (figure 7.8 (c)). It could be caused by not finding a gene in this region by a program or basically by none of the gene existing within the particular region. It was found that UTR5 is the most deleted region by appearing 151 times, followed by CDS with 136 events and UTR3 with only 118. The numbers of all deleted events are shown in figure 7.8 (b) where the regions are counted by annotated regions for each deletion site. Thus the CDS means that 38 deletion sites have deleted only CDS region and none of the UTRs. The lowest count is for deletion sites that deleted both UTR5 and UTR3 without deleting CDS region. Furthermore, the program identified 251 genes corresponding to 215 deletion sites labeled as 'dispensable' genes.

At last, the minimum RVIS score for all genes within a deletion site was identified. If the deletion site with annotated genes does not have a RVIS score, it means that there is no record of RVIS score for these genes. The mean RVIS of all sites was 0.76, implying that these homozygously deleted genes are highly tolerant of mutation.

Comparison with the 1000 Genomes Project results

The results are compared to the reference results produced by the 1000 Genomes Project downloaded from a supplementary material of phase 3 SV release *paper* named `Table_S6`. In general, the 1000 Genomes Project has identified 240 genes corresponding to 204 individual deletion sites that were affected by homozygous losses and named as 'dispensable' genes. In contrast, the project analysis identified 251 genes corresponding to 215 sites. Next, deleted regions discovered by the 1000 Genomes Project are compared with analysis results within the figure 7.9. The last two groups 'CDS' and 'UTR' show numbers of these regions regardless of another region appearing within an individual site. Almost every group showed the excess in analysis results except the 'only CDS' group with a majority of the 1000 Genomes project results. The biggest excess can be seen within the 'UTR' group (200 against 155). In contrast, 'CDS' group has the most similar values between analysis and the 1000 Genomes Project results, 38 and 49 respectively. The differences could be caused by downloading different dataset of CDS and UTR regions from the UCSC table or by the fact that the database could have changed over the years.

Furthermore, the figure 7.10 presents RVIS score within analysis and reference results. The mean in both was set to 0.76 showing that these homozygously deleted genes are tolerant of mutation. The median in both results is the same and the minimum value cutoff is on bigger values within reference results, regardless potential outliers. Also, the 25th percentile is a bit bigger in reference results.

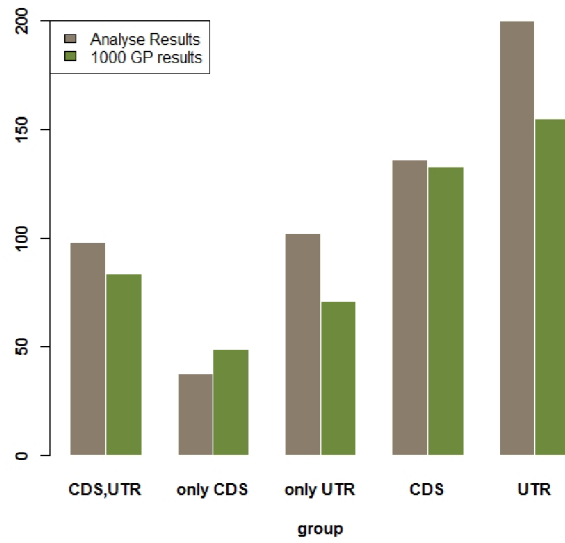


Fig. 7.9: Deleted regions found by analysis compared with a reference numbers from the 1000 Genomes Project.

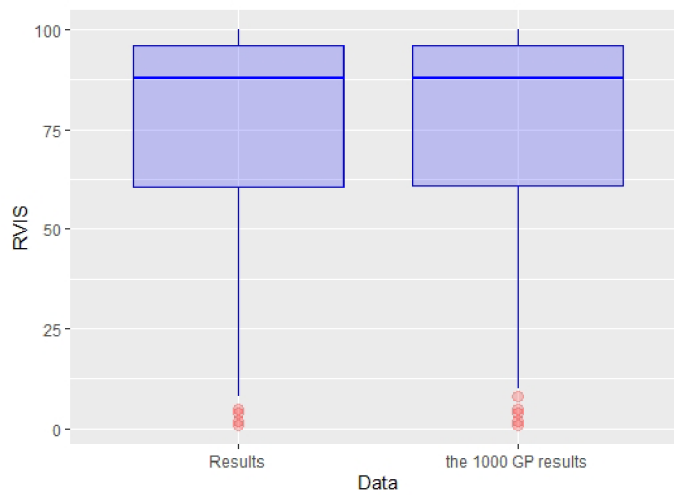


Fig. 7.10: RVIS score within analysis results and the 1000 Genomes Project results. The RVIS score is in %. Red dots are showing potential outliers and the straight blue lines within boxplots represent median. The vertical line shows the minimum and maximum values in dataset on its endings.

Conclusion

The bachelor's thesis aimed to study the human genome variations which were found by the HapMap Project and the 1000 Genomes Project. Each project has several phases. More than 1 million SNPs were genotypes after a phase of the HapMap Project. The second phase has extended the database by adding 2.1 million novel SNPs. These two phases discovered most of the common SNPs but the progress needs to be done in rare variants discovery. Thus the HapMap 3 aimed to genotype rare variants and also copy number polymorphisms (CNPs) with an extended sample set. The HapMap Project made disease research faster and cheaper by genotyping 500,000 tag SNPs using a linkage disequilibrium instead of 10 million. The 1000 Genomes Project has overcome the number of SNPs discovered by the HapMap Project. The pilot phase discovered around 20 million SNPs, 2 million INDELS and 34,000 structural variants. Phase 1 identified 38 million SNPs, 1.4 million short indels and over 14,000 larger deletions. During phase 2, the number of samples increased and methods were improved. At last, around 88 million variants were identified during phase 3 with the addition of 68,818 SVs discovered by its second part.

Further, both projects databases were explained and The Genomes Browser was annotated as an online tool allowing to browse and download their data.

The second part of the bachelor's thesis analysed data from the phase 3 SV discovery released by the 1000 Genomes Project. Population, statistical and bioinformatic analyses were run on data. PCA showed that the African populations differ the most among the other populations which are clustering close together. Next, the biggest number of heterozygous events was exhibited again by the African populations, corresponding to the increased diversity of their individuals. Vst showed that only a small portion of deletions, duplications and mCNV are highly-stratified and likely of undergoing adaptive selection. Bioinformatic analysis identified 251 'dispensable' genes corresponding to 215 deletion sites that removed at least one entire exon, and RVIS score along these genes showed that they are tolerant of mutation.

Bibliography

- [1] About IGSR and the 1000 Genomes Project. *IGSR: The International Genome Sample Resource* [online]. [cit. 2019-11-23]. URL: <https://www.internationalgenome.org/about>
- [2] ALKAN, Can, Bradley P. COE and Evan E. EICHLER. Genome structural variation discovery and genotyping. *Nature Reviews Genetics* [online]. 2011, **12**(5), 363-376 [cit. 2019-12-29]. DOI: 10.1038/nrg2958. ISSN 1471-0056. URL: <http://www.nature.com/articles/nrg2958>
- [3] ALTSHULER, D., P. DONNELLY. A haplotype map of the human genome. *Nature* [online]. 2005, **437**(7063), 1299-1320 [cit. 2019-11-17]. DOI: 10.1038/nature04226. ISSN 0028-0836.
- [4] ALTSHULER, D., R. GIBBS, L. PELTONEN, et al. Integrating common and rare genetic variation in diverse human populations. *Nature* [online]. 2010, **467**(7311), 52-58 [cit. 2019-11-19]. DOI: 10.1038/nature09298. ISSN 0028-0836. URL: <http://www.nature.com/articles/nature09298>
- [5] ANSORGE, Wilhelm J. Next-generation DNA sequencing techniques. *New Biotechnology*. 2009, **25**(4), 195-203. DOI: 10.1016/j.nbt.2008.12.009. ISSN 18716784. URL:
- [6] AUTON, A., g. ABECASIS, D. ALTSHULER, et al. A global reference for human genetic variation. *Nature* [online]. 2015, **526**(7571), 68-74 [cit. 2019-11-25]. DOI: 10.1038/nature15393. ISSN 0028-0836.
- [7] BUCHANAN, Carrie C., Eric S. TORSTENSON, William S. BUSH and Marylyn D. RITCHIE. A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. *Journal of the American Medical Informatics Association* [online]. 2012, **19**(2), 289-294 [cit. 2019-12-09]. DOI: 10.1136/amiajnl-2011-000652. ISSN 1067-5027.
- [8] CLIFTEN, Paul, A. AUTON, G. ABECASIS, et al. Base Calling, Read Mapping, and Coverage Analysis. *Clinical Genomics*. Elsevier, 2015, 2015, **27**(15), 91-107. DOI: 10.1016/B978-0-12-404748-8.00007-1. ISBN 9780124047488. ISSN 1367-4803.
- [9] COMMINS, Jennifer, Christina TOFT and Mario A. FARES. Computational Biology Methods and Their Application to the Comparative Genomics of Endocellular Symbiotic Bacteria of Insects. *Biological Procedures Online*.

- 2009, **11**(1), 52-78. DOI: 10.1007/s12575-009-9004-1. ISSN 1480-9222. URL: <http://www.biologicalproceduresonline.com/content/11/1/52>
- [10] CVRČKOVÁ, Fatima. *Úvod do praktické bioinformatiky*. Praha: Academia, 2006. ISBN 80-200-1360-1.
- [11] DALMA-WEISZHAUSZ, Dennise D., Janet WARRINGTON, Eugene Y. TANIMOTO and C. Garrett MIYADA. [1] The Affymetrix GeneChip® Platform: An Overview. *DNA Microarrays, Part A: Array Platforms and Wet-Bench Protocols*. Elsevier, 2006, 2006, **3**(2), 3-28. Methods in Enzymology. DOI: 10.1016/S0076-6879(06)10001-4. ISBN 9780121828158. ISSN 17406749. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0076687906100014>
- [12] DANECEK, P., A. AUTON, G. ABECASIS, et al. The variant call format and VCFtools. *Bioinformatics*. 2011, **27**(15), 2156-2158. DOI: 10.1093/bioinformatics/btr330. ISSN 1367-4803.
- [13] DEFUSCO, Richard ARMAND, Dennis W. MCLEAVEY, Jerald E. PINTO and David E. RUNKLE. *Quantitative investment analysis*. Third edition. Hoboken, New Jersey: Wiley, [2015]. ISBN 978-111-9104-223.
- [14] DEVUYST, O. The 1000 Genomes Project: Welcome to a New World. *Peritoneal Dialysis International* [online]. 2015, **35**(7), 676-677 [cit. 2019-11-22]. DOI: 10.3747/pdi.2015.00261. ISSN 0896-8608.
- [15] DEWEERDT, S. E. *What's a Genome?* [online]. January 2003 [cit. 2019-11-08].
- [16] DUAN, Shiwei, Wei ZHANG, Nancy Jean COX and Mary Eileen DOLAN. FstSNP-HapMap3: a database of SNPs with high population differentiation for HapMap3. *Bioinformation* [online]. 2008, **3**(3), 139-141 [cit. 2019-11-19]. DOI: 10.6026/97320630003139. ISSN 09738894.
- [17] DURBIN, R., D. ALTSHULER, R. DURBIN, et al. A map of human genome variation from population-scale sequencing. *Nature* [online]. 2010, **467**(7319), 1061-1073 [cit. 2019-11-22]. DOI: 10.1038/nature09534. ISSN 0028-0836. URL: <http://www.nature.com/articles/nature09534>
- [18] FEUK, Lars, Andrew R. CARSON and Stephen W. SCHERER. Structural variation in the human genome. *Nature Reviews Genetics* [online]. 2006, **7**(2), 85-97 [cit. 2019-12-29]. DOI: 10.1038/nrg1767. ISSN 1471-0056. URL: <http://www.nature.com/articles/nrg1767>

- [19] FRAZER, K., D. BALLINGE, D. COX, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* [online]. 2007, **449**(7164), 851-861 [cit. 2019-11-18]. DOI: 10.1038/nature06258. ISSN 0028-0836. URL: <http://www.nature.com/articles/nature06258>
- [20] FRIDOVICH-KEIL, Judith L. Human genome. *Encyclopædia Britannica* [online]. Chicago: Encyclopædia Britannica, 15 February 2019 [cit. 2019-12-28]. URL: <https://www.britannica.com/science/human-genome>
- [21] *Genic Intolerance* [online]. New York: Institute for Genomic Medicine [cit. 2020-05-24]. URL: <http://genic-intolerance.org>
- [22] GIBBS, R., J. BELMONT, P. HARDENBOL, et al. The International HapMap Project. *Nature* [online]. 2003, **426**(6968), 789-796 [cit. 2019-11-16]. DOI: 10.1038/nature02168. ISSN 0028-0836.
- [23] GRIFFITHS, Anthony JF, Jeffrey H. MILLER, David T. SUZUKI, Richard C. LEWONTIN and William M. GELBART. *An Introduction to Genetic Analysis*. 7th edition. New York: W. H. Freeman, 2000. ISBN 0-7167-3520-2.
- [24] *International HapMap Project FTP-repository* [online]. [cit. 2020-05-01]. URL: <ftp://ftp.ncbi.nlm.nih.gov/hapmap/>
- [25] International HapMap Project. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001-, 16 September 2019 [cit. 2019-11-18].
- [26] JOLLIFFE, Ian T. and Jorge CADIMA. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2016, **374**(2065). DOI: 10.1098/rsta.2015.0202. ISSN 1364-503X.
- [27] KIM, Sobin. SNP genotyping: technologies and biomedical applications. *Annual Review of Biomedical Engineering*. 2007, **9**, 289—320. DOI: 10.1146/annurev.bioeng.9.060906.152037.
- [28] KULKARNI, Shashikant and John D. PFEIFER, ed. *Clinical genomics*. Amsterdam: Elsevier, c2015. ISBN 978-0-12-404748-8.
- [29] MAYER, Melissa. Difference Between Homozygous Heterozygous. *Sciencing.com* [online]. [cit. 2020-05-23]. URL: <https://sciencing.com/difference-between-homozygous-heterozygous-8606730.html>

- [30] McVEAN, G., D. ALTSHULER (Co-Chair), R. DURBIN, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* [online]. 2012, **491**(7422), 56-65 [cit. 2019-11-23]. DOI: 10.1038/nature11632. ISSN 0028-0836. URL: <http://www.nature.com/articles/nature11632>
- [31] PAREEK, Chandra SHEKHAR, Rafal SMOCZYNSKY, Andrzej TRETYN and Luan FEI-SHI. Sequencing technologies and genome sequencing. *Journal of Applied Genetics*. 2011, **52**(4), 413-435. DOI: 10.1007/s13353-011-0057-x. ISSN 1234-1983. URL: <http://link.springer.com/10.1007/s13353-011-0057-x>
- [32] QIANG-LONG, Zhu, Liu SHI, Gao PENG and Luan FEI-SHI. High-throughput Sequencing Technology and Its Application. *Journal of Northeast Agricultural University (English Edition)*. 2014, **21**(3), 84-96. DOI: 10.1016/S1006-8104(14)60073-8. ISSN 10068104.
- [33] RAGOISSIS, Jiannis. Genotyping Technologies for Genetic Research. *Annual Review of Genomics and Human Genetics*. 2009, **10**, 117-133. DOI: 10.1146/annurev-genom-082908-150116.
- [34] RAGOISSIS, Jiannis. Genotyping technologies for all. *Drug Discovery Today: Technologies*. 2006, **3**(2), 115-122. DOI: 10.1016/j.ddtec.2006.06.013. ISSN 17406749.
- [35] REUTER, Jason A., Damek V. SPACEK, Michael P. SNYDER and Luan FEI-SHI. High-Throughput Sequencing Technologies. *Molecular Cell*. 2015, **58**(4), 586-597. DOI: 10.1016/j.molcel.2015.05.004. ISSN 10972765. URL: <https://linkinghub.elsevier.com/retrieve/pii/S1097276515003408>
- [36] REICH, David, Alkes L PRICE and Nick PATTERSON. Principal component analysis of genetic data. *Nature Genetics*. 2008, **40**(5), 491-492. DOI: 10.1038/ng0508-491. ISSN 1061-4036.
- [37] RHOADS, Anthony and Kin Fai AU. *PacBio Sequencing and Its Applications*. 2015, **13**(5), 278-289. DOI: 10.1016/j.gpb.2015.08.002. ISSN 16720229.
- [38] ROTH, Stephen M. *Genetics primer for exercise science and health*. Champaign, IL: Human Kinetics, c2007. ISBN 07-360-6343-9.
- [39] SMITH, Lindsay I. *A tutorial on Principal Components Analysis*. 2002. URL: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf
- [40] SUDMANT, Peter H., Tobias RAUSCH, Eugene J. GARDNER, et al. An integrated map of structural variation in 2,504 human genomes. *Nature* [online].

- 2015, **526**(7571), 75—81 [cit. 2019-11-26]. DOI: 10.1038/nature15394. ISSN 0028-0836. URL: <https://www.nature.com/articles/nature15394#MOESM91>
- [41] 1000 Genomes Browser. *NCBI* [online]. Bethesda (MD): National Center for Biotechnology Information, U.S. National Library of Medicine [cit. 2019-12-05]. URL: <https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/help/>
- [42] WESTHEAD, David R., J. Howard PARISH and Richard M. TWYMAN. *Instant Notes in Bioinformatics*. Oxford: BIOS, 2002. ISBN 1859962726.
- [43] What is clone-by-clone sequencing? *Yourgenome* [online]. 2017 [cit. 2020-05-27]. URL: <https://www.yourgenome.org/facts/what-is-clone-by-clone-sequencing>
- [44] YUAN, Yuan, Lei TIAN, Dongsheng LU and Shuhua XU. Analysis of Genome-Wide RNA-Sequencing Data Suggests Age of the CEPH/Utah (CEU) Lymphoblastoid Cell Lines Systematically Biases Gene Expression Profiles. *Scientific Reports*. 2015, **5**(1). DOI: 10.1038/srep07960. ISSN 2045-2322. URL: <http://www.nature.com/articles/srep07960>


A 1000 Genomes Project Populations

Tab. A.1: List of populations with annotated continental groups from the 1000 Genomes Project. I stands for phase I, II for phase II and III for phase III. Population with annotated phase III were used for analyses. [17] [30] [40]

Population	Continental Group	Phase
African Ancestry in Southwest US (ASW)	African (AFR)	I/III
African Caribbean in Barbados (ACB)	African (AFR)	III
Bengali in Bangladesh (BEB)	South Asian (SAS)	III
British in England and Scotland (GBR)	European (EUR)	I/III
Chinese in Denver, Colorado (USA)	East Asian (EAS)	Pilot
Chinese Dai in Xishuangbanna, China (CDX)	East Asian (EAS)	III
Colombian in Medellin, Colombia (CLM)	American (AMR)	I/III
Esan in Nigeria (ESN)	African (AFR)	III
Finnish in Finland (FIN)	European (EUR)	I/III
Gambian in Western Division, The Gambia (GWD)	African (AFR)	III
Gujarati Indian in Houston, TX (GIH)	South Asian (SAS)	III
Han Chinese in Beijing, China (CHB)	East Asian (EAS)	Pilot/I/III
Iberian populations in Spain (IBS)	European (EUR)	I/III
Indian Telugu in the UK (ITU)	South Asian (SAS)	III
Japanese in Tokyo, Japan (JPT)	East Asian (EAS)	Pilot/I/III
Kinh in Ho Chi Minh City, Vietnam (KHV)	East Asian (EAS)	III
Luhya in Webuye, Kenya (LWK)	African (AFR)	Pilot/I/III
Mende in Sierra Leone (MSL)	African (AFR)	III
Mexican Ancestry in Los Angeles, California (MXL)	American (AMR)	I/III
Peruvian in Lima, Peru (PEL)	American (AMR)	III
Puerto Rican in Puerto Rico (PUR)	American (AMR)	I/III
Punjabi in Lahore, Pakistan (PJL)	South Asian (SAS)	III
Southern Han Chinese, China (CHS)	East Asian (EAS)	I/III
Sri Lankan Tamil in the UK (STU)	South Asian (SAS)	III
Toscani in Italy (TSI)	European (EUR)	Pilot/I/III
Utah residents with European ancestry (CEU)	European (EUR)	Pilot/I/III
Yoruba in Ibadan, Nigeria (YRI)	African (AFR)	Pilot//I/III

B Downloading VCF file

Index of /vol1/ftp/phase3/integrated_sv_map/

 [parent directory]







	Name	Size	Date Modified
	ALL.autosomes.pindel.20130502.complexindex.low_coverage.genotypes.vcf.gz	396 kB	1/31/15, 1:00:00 AM
	ALL.autosomes.pindel.20130502.complexindex.low_coverage.genotypes.vcf.gz.tbi	13.3 kB	1/31/15, 1:00:00 AM
	ALL.wgs.mergedSV.v8.20130502.svs.genotypes.vcf.gz	17.5 MB	5/19/17, 2:00:00 AM
	ALL.wgs.mergedSV.v8.20130502.svs.genotypes.vcf.gz.tbi	651 kB	5/19/17, 2:00:00 AM
	README_phase3_sv_callset_20150224	5.8 kB	5/21/17, 2:00:00 AM
	supporting/		1/31/15, 1:00:00 AM

Fig. B.1: The 1000 Genomes Project FTP site database containing phase 3 SV release files.

The figure B.1 shows the *1000 Genomes FTP Site* with the 1000 Genomes Project phase 3 Structural Variant discovery release. A path to this subdirectory from the main page is `phase3/ ->integrated_sv_map/`. Via this *link* the `ALL.wgs.mergedSV.v8.20130502.svs.genotypes.vcf.gz` file containing the structural variant data will be automatically downloaded. The downloaded file is zipped in gzip archive. The file can be unzip via WinRAR, 7-Zip or via an *online tool*. The unzip vcf file is downloaded into R environment by using `vcfR` package. A vcf file is uploaded as `read.vcfR('vcf file needs to be added here')`.

C Downloading CDS and UTR regions from UCSC Table Browser

To download CDS, UTR5 and UTR3 regions from the *UCSC Table Browser* it is necessary to change default setting of the page to be the same as in figure C.1 [a]. Output format is set to BED and in output file column it is necessary to write its name with `bedfile` extension. The regions are downloaded separately, thus the output file column should be filled with either `cds.bed`, `utr5.bed` or `utr3.bed`. After `get output` button is pushed, the page in figure C.1 [b] will pop up. On this page, it is possible to choose which region will be listed. The chosen region types for this analysis are : Coding Exons, 5'UTR Exons and 3'UTR Exons to get CDS, UTR5 and UTR3 regions respectively. The final BED file with regions positions is downloaded through the `get BED` button. The BED file is constituted with chromosome name, start and end positions, and annotated genes for particular regions. BED file is then uploaded into R environment through the `findRegions` function within `gene.R` script.

a

clade: Mammal genome: Human assembly: Feb. 2009 (GRCh37/hg19)
group: Genes and Gene Predictions track: NCBI RefSeq
table: RefSeq All (ncbiRefSeq)
region: genome ENCODE Pilot regions position
identifiers (names/accessions):
filter: create
subtrack merge: create
intersection: create
correlation: create
output format: BED - browser extensible data Send output to Galaxy GREAT
output file: cds.bed (leave blank to keep output in browser)
file type returned: plain text gzip compressed
get output summary/statistics

b

Create one BED record per:

- Whole Gene
- Upstream by bases
- Exons plus bases at each end
- Introns plus bases at each end
- 5' UTR Exons
- Coding Exons
- 3' UTR Exons
- Downstream by bases

Note: if a feature is close to the beginning or end of a chromosome, you may want to use the "Upstream" or "Downstream" options to avoid extending past the edge of the chromosome.

get BED cancel

Fig. C.1: UCSC Table Browser pages